TECHNISCHE UNIVERSITÄT WIEN

DISSERTATION

Spatial dependence, trends, functional outliers and sparsity in Compositional Data Analysis

ausgeführt zum Zwecke der Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften unter der Leitung von

Peter Filzmoser
Institute of Statistics and Mathematical Methods in Economics, TU Wien

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

Christopher Rieser
Matrikelnummer: 00815803

Diese Dissertation haben begutachtet:

1. **Univ.Prof. Dr.techn. Peter Filzmoser**
   Institute of Statistics and Mathematical Methods in Economics, TU Wien
2. **Prof. RNDr. Karel Hron PhD**
   Department of Mathematical Analysis and Applications of Mathematics, Palacky University
3. **Prof. Anne Ruiz-Gazen PhD**
   TSE Research Faculty, University of Toulouse Capitole

Wien, am 8. Februar 2022

Abstract

Many real world datasets are nowadays multivariate. The last decades have seen an explosion of the number of methods for multivariate data. Such data, usually considered to be real valued, and the corresponding statistical methods, are based on the Euclidean geometry. The underlying property of the Euclidean geometry is to measure similarity of two points as the length of the line between such. For settings in which data points consist of strictly positive variables, which are plentiful in the applied sciences, a different approach needs to be considered. Recently, a concept quickly gaining popularity, called Compositional Data Analysis (CoDA) has emerged. In CoDA one measures similarity of two multivariate strictly positive data points by comparing the log-ratios respectively. With this change of perspective comes the necessity to adapt classical statistical methods. Many such methods have been extended to this framework in recent years. In this thesis we look into applications of a combined approach between the log-ratio methodology, which is at the heart of CoDA, and Generalized Additive Models, to find important elements in a geochemical exploration setting. Furthermore we explore various extensions of CoDA such as change point detection of compositional time series, outlier detection of compositional functional data and very fruitful connections with signal processing on graphs.

Kurzfassung

Viele reale Datensätze sind zum gegenwärtigen Zeitpunkt multivariat. In den letzten Jahrzehnten ist die Zahl der Methoden für multivariate Daten geradezu explodiert. Daten, die normalerweise als reell angesehen werden, basieren jedoch auf entsprechenden statistischen Methoden die wiederum auf die euklidische Geometrie zurückzuführen sind. Die zugrundeliegende Annahme der euklidischen Geometrie besteht darin, den Abstand zweier Punkte als die Länge der Linie zwischen diesen zu definieren. In Fällen, in denen Datenpunkte aus strikt positiven Variablen bestehen, diese Beispiele findet man in den angewandten Wissenschaften reichlich, muss ein anderer Ansatz in Betracht gezogen werden. In der jüngsten Vergangenheit gewinnt ein Konzept namens Compositional Data Analysis (CoDA) immer mehr an Popularität. In CoDA misst man den Abstand zweier mulivariater strikt positiver Datenpunkte, indem man die entsprechend jeweiligen logarithmisch transformierten Quotienten aller Variablen vergleicht. Dieser Perspektivenwechsel führt zur Notwendigkeit, die klassischen statistischen Methoden anzupassen. Zahlreiche klassische Methoden der letzten Jahre wurden an solche Daten angepasst. In dieser Dissertation untersuchen wir Anwendungen eines kombinierten Ansatzes zwischen der Log-Ratio-Methodik, welche das Herzstück von CoDA ist, und generalisierten additiven Modellen, um interessante Elemente einer geochemischen Exploration zu finden. Darüber hinaus erforschen wir verschiedene Erweiterungen von CoDA, wie die Erkennung von Änderungspunkten aus Kompositionszeitreihen, der Erkennung von Ausreißern von funktionellen Kompositionsdaten und außerordentlich fruchtbaren Verbindungen mit der Signalverarbeitung auf Graphen.

Acknowledgement

This page is, in my eyes, the most important page of this thesis as it gives me a chance to say thank you, even if it is just a few words. I would not have gotten so far without the people who supported me, may it be in academic questions or personal ones. I've learned these last few years that success in research is only possible with the support of one's colleagues and, not exclusively of course, friends and family.

I am very fortunate to have met Peter, my supervisor, at such an early stage of my career. His knowledgeable input, patience when having long discussions as well as patience whenever I thought I had come up with a good, but in reality chaotic, idea, was never-ending. His openness to talk about any problem, academic or not, was infinite. For these reasons, I do not only respect him very much as an academic but also as a person. Long story short, thank you a thousand times!

Another thousand thank-yous go to my partner and pandemic-related kitchen office colleague, Marlene. Whenever I spent long hours working on my doctorate or on an idea that had caught my attention, Marlene was always very understanding and supportive. She made me laugh whenever I was on the brink of mathematically induced insanity, and this made me appreciate her even more. I couldn't have done it without her and for this reason, I will always be grateful.

Again, a thousand thank-yous go, how could it be otherwise, to my mother. Thank you for the emotional support and the advice you have given me coming from your own PhD experience. I consider you my role model. Thank you!

An additional thousand thank-yous go to the two people who have agreed to be my referees, Karel Hron and Anne Ruiz-Gazen. I greatly appreciate the time you took to read through my thesis! Thank you. I will also always remember the week I spent in Olmouc.

A special nine hundred ninety-nine thank-yous go out to Christoph. I count myself very lucky to have had the chance to share the office and many laughs with such a great colleague and friend. I also want to thank my two brothers, Patrick and Daniel, as well as my friends Sipo, Verena, Dan, Jerome, Verena, Alex, Kathi, Camilla together with olomoucké syrečky, Domi, Ana and all my work colleagues. Finally, even though they do not understand human language, thank you Funky, Milka and Lady.

Thinking of the time of my doctorate will always bring a smile to my face and the memory of a little sleep deprivation back to my mind.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am 8. Februar 2022

Christopher Rieser

Original publications

This thesis consists of two parts, a general introduction to recent developments of Compositional Data Analysis and the following publications:

I D. Mikšová, C. Rieser, and P. Filzmoser. Identification of mineralization in geochemistry along a transect based on the spatial curvature of log-ratios. Mathematical Geosciences 53, 1513-1533, 2021. https://doi.org/10.1007/s11004-021-09930-4.

II D. Mikšová, C. Rieser, P. Filzmoser, M. Middleton, and R. Sutinen. Identification of mineralization in geochemistry for grid sampling using Generalized Additive Models. Mathematical Geosciences 53, 1861–1880, 2021. https://doi.org/10.1007/s11004-021-09929-x.

III C. Rieser, P. Filzmoser. Compositional trend filtering. Annales Mathematicae et Informaticae, 53, 257-270, 2021. https://doi.org/10.33039/ami.2021.02.004.

IV C. Rieser, P. Filzmoser. Outlier Detection for Pandemic-Related Data Using Compositional Functional Data Analysis. In: Boado-Penas M.C., Eisenberg J., Sahin S (eds). Pandemics: Insurance and Social Protection. Springer Actuarial, Springer, Cham, 2022. https://doi.org/10.1007/978-3-030-78334-1_12.

V C. Rieser, P. Filzmoser. Extending compositional data analysis from a graph signal processing perspective. Submitted for publication.

Contents

Part I.

Summary

1. Introduction

One of the first steps when starting to analyze a dataset is to look at the type of data one has and consequently which sample space one should use. The choice of methods used for real valued, categorical, time series, strictly positive, interval, tensor, functional or image data, and many more, can have a huge impact on the outcome of the analysis. Recent decades have seen a surge in methods that allows one to incorporate any such knowledge into the analysis.

Compositional Data Analysis (CoDA) deals with multivariate, strictly positive data whose components are part of a whole. This loose first definition of CoDA happens to be fulfilled for many real world datasets we are interested in. A very common example are household expenditures of different EU countries for categories like food, housing, health, leisure, transport, etc. Each country represents a multivariate datapoint with strictly positive entries which can be thought of as percentages in regard to the sum of their expenditures. Commonly used approaches to such data as to divide each entry by the sum of entries and then log-transform respectively, ignore spurious correlation effects, see [48], and lead to false conclusions. Latter is one of the main motivations for the development of CoDA, such as in [1] and [3]. Working in the right (compositional) geometry incorporates for these examples the viewpoint of percentages and facilitates any further analysis greatly. Further examples are omnipresent, such as microbiome data [23] or [51], omics data [49] or time-use data [16]. However, despite all the work that has gone into this field many applied areas are still unaware of the advantages [40].

Obviously, different extensions of CoDA arise in case that we also want to include additional information in an analysis. Consider a setting in which we are interested in analyzing expenditures of different households in a given city. A natural question that arises is how strong expenditures vary from one part of the city to another one or if certain types of expenditures dominate in certain areas. Modeling such geographic dependencies is commonly done in geostatistics, however methods need to be carefully chosen and adapted in case a compositional setting is chosen. Additionally to geospatial information, we might be interested in how expenditures vary over a given period of time and if behaviors and the general trend changes at some point. Classically, time series are considered to be real valued. A CoDA approach, on the other hand, considers a different point of view than in the classical case. At last, assume we are interested to find EU countries that display an outlying behavior from the average one over time in terms of their expenditures. Such a task could be modeled by a compostionally adapted approach for functional outlier detection.

Evidently, there are plenty extensions of CoDA which are interesting. In this thesis we consider variations of the mentioned ones. The structure is as follows. Section 2 gives, after a short example, an overview of the most important concepts of CoDA, and introduces all necessary tools and core concepts. In the subsequent sections we give a compact overview of the main scientific contributions as developed in Papers I to V, which include the use of Generalized Additive Models phrased in a compositional context for geochemical anomaly detection, compositional time series trend detection, compositional Functional Data outlier

detection and a rephrasing of CoDA in the context of Graph Signal Processing. In the last section we look at possible future research directions and give an outlook.

2. Basic concepts of CoDA

The following subsections introduce core concepts of CoDA. We will start with a subsection that motivates the principles and definitions which follow.

2.1. Core philosophy of CoDA

To illustrate core principles of CoDA we look at an example of household expenditures, as mentioned in the introduction, first. The data is available in the the R package robCompositions [58] and consists of twelve different types of expenditures - such as Food, Alcohol, Clothing, Housing, Furnishings, Health, Transport, Communications, Recreation, Education, Restaurants and Other - from households in 27 EU countries. Table 2.1 shows an excerpt of expenditures in Euros for the countries Denmark (DK), Germany (D), Irland (IRL), Spain (ES) and France (F). If we are interested in how much an average Irish household spends on food compared to a Spanish one we see, looking at the absolute numbers, that in both countries food expenditures are similar - 4491 Euro for Ireland and 4685 Euro for Spain. Something similar holds for a German and Irish household in the housing category - 8445 Euro for Germany and 8520 Euro for Ireland. However, since in each country the average total spending budget of a household differs due to different economic factors it is questionable that such a comparison provides a fair picture. It can be argued that it is more sensible to compare relative numbers, that is comparing the percentages an average household of a country spends in each country compared to their total expenditures. Table 2.2 shows these percentages and the picture is quite different. The average Spanish household spends 18% of their total budget on food whereas an average Irish household only spends 12.35 % on food. As mentioned in the introduction, we might think of CoDA, in a first approach, as percentages. Percentages have two main properties.

	Food	*Alcohol*	*Clothing*	*Housing*	*Furnishings*	*Health*	*Transport*	*Communications*	*Recreation*	*Education*	*Restaurants*	*Other*
DK	2872	785	1168	7194	1459	639	3331	583	2738	100	960	2233
D	3185	489	1355	8445	1543	1024	3790	828	3168	236	1212	3226
IRL	4491	2032	1851	8520	2613	904	4203	1255	3670	687	2190	3956
ES	4685	586	1786	7874	1211	577	2743	701	1659	292	2414	1499
F	3733	650	1853	7339	1693	1167	3777	914	1926	165	1277	3392

Table 2.1.: Different types of expenditures in Euros for five EU countries.

	Food	*Alcohol*	*Clothing*	*Housing*	*Furnishings*	*Health*	*Transport*	*Communications*	*Recreation*	*Education*	*Restaurants*	*Other*
DK	11.94	3.26	4.85	29.90	6.06	2.66	13.84	2.42	11.38	0.42	3.99	9.28
D	11.18	1.72	4.75	29.63	5.41	3.59	13.30	2.91	11.12	0.83	4.25	11.32
IRL	12.35	5.59	5.09	23.42	7.18	2.49	11.56	3.45	10.09	1.89	6.02	10.88
ES	18.00	2.25	6.86	30.25	4.65	2.22	10.54	2.69	6.37	1.12	9.27	5.76
F	13.39	2.33	6.64	26.32	6.07	4.18	13.54	3.28	6.91	0.59	4.58	12.16

Table 2.2.: Different types of expenditures in Euros for five EU countries in percentages of the sum of expenditures.

The first property is that they are always positive. The second property is that they always sum to one. In the expenditures example this is enforced by dividing each expenditure by the sum of expenditures. Such an operation is invariant under rescaling, that is, if we multiply, for a fixed country, each expenditure by a strictly positive constant, $\lambda > 0$, then the percentages would not change. We define the space of D-dimensional positive reals as $\mathbb{R}_+^D := \{(x_1, \ldots, x_D)' \in \mathbb{R}^D \mid x_j > 0 \forall j = 1, \ldots, D\}$. For two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}_+^D$ that fulfill $\boldsymbol{y} = \lambda \boldsymbol{x}$ we get that their scaled versions, i.e. percentages, are equal, shown by the simple calculation

$$\frac{1}{\sum_{i=1}^D y_i} \boldsymbol{y} = \frac{1}{\sum_{i=1}^D \lambda x_i} \lambda \boldsymbol{x} = \frac{\lambda}{\lambda \sum_{i=1}^D x_i} \boldsymbol{x} = \frac{1}{\sum_{i=1}^D x_i} \boldsymbol{x}.$$

The second property can therefore be rephrased as regarding $\boldsymbol{x}$ and $\boldsymbol{y}$ as equivalent, written as $\boldsymbol{x} \sim \boldsymbol{y}$, if and only if a $\lambda > 0$ exists such that $\boldsymbol{y} = \lambda \boldsymbol{x}$ holds.

Figure 2.1 shows for different fixed $\boldsymbol{x}$ the lines $\{\lambda \boldsymbol{x} | \lambda > 0\}$ starting in zero and extending to infinity for the case of $D = 2$. Each line starting in zero and passing through $\boldsymbol{x}$ can be identified with the point $(\sum_{i=1}^D x_i)^{-1} \boldsymbol{x}$. The set of points $\{\sum_{i=1}^D x_i)^{-1} \boldsymbol{x} | \boldsymbol{x} \in \mathbb{R}_+^D\}$ is shown as a line from $(0, 1)$ to $(1, 0)$. To be able to use most statistical tools a vector space structure, a distance measure and sometimes an inner product is a prerequisite. Any such definitions need to be invariant under a chosen point on a line as depicted in Figure 2.1. It is easy to see that applying log entrywise to any element $\boldsymbol{x} \in \mathbb{R}_+^D$ gives the whole of $\mathbb{R}^D$. As $\mathbb{R}^D$ is a vector space with componentwise addition and scalar multiplication we could turn $\mathbb{R}_+^D$ into a vector space by taking the inverse of the mapping $\log : \mathbb{R}_+^D \to \mathbb{R}^D$. Defining a norm on $\mathbb{R}_+^D$ in the same way does however not lead to an invariant norm and thus this approach can not be used. An appropriate approach to such data was laid out in [3] and [2] leading to the monograph [4]. This is discussed in the following sections.

2.2. A (very) quick introduction to finite dimensional real Hilbert spaces

Before defining main principles of the Aitchison geometry we want look at the core concepts of finite dimensional real Hilbert spaces.

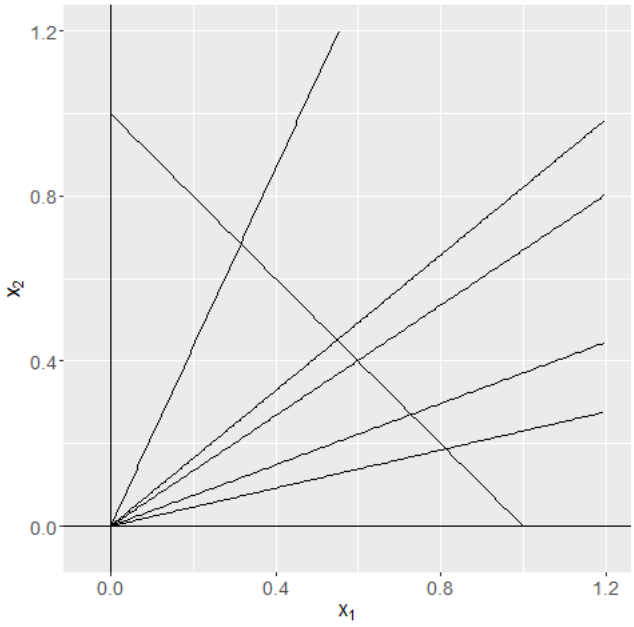

Fig. 2.1.: Different lines $\{\lambda \boldsymbol{x} | \lambda > 0\}$ starting in zero and extending to infinity. Each line can be identified with a point on the line from $(0,1)$ to $(1,0)$.

Definition 1. *A finite dimensional real Hilbert space $(\mathcal{H}, +, \cdot, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ is a finite dimensional real vector space $\mathcal{H}$ equipped with an addition $+$ and scalar multiplication operation $\cdot$, see [52], and a function $\langle\cdot,\cdot\rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$, also called an inner product, such that for any $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathcal{H}$ and $\alpha, \beta \in \mathbb{R}$ we have*

- *linearity in the second argument*

$$\langle \boldsymbol{u}, \alpha\boldsymbol{v} + \beta\boldsymbol{w}\rangle_{\mathcal{H}} = \langle \boldsymbol{u}, \alpha\boldsymbol{v}\rangle_{\mathcal{H}} + \langle \boldsymbol{u}, \beta\boldsymbol{w}\rangle_{\mathcal{H}},$$

- *symmetry, $\langle \boldsymbol{u}, \boldsymbol{v}\rangle_{\mathcal{H}} = \langle \boldsymbol{v}, \boldsymbol{u}\rangle_{\mathcal{H}}$,*
- *and positivity $\langle \boldsymbol{u}, \boldsymbol{u}\rangle_{\mathcal{H}} \geq 0$, with equality only when $\boldsymbol{u} = 0$.*

The definition of an inner product on a Hilbert space is motivated by the inner product on the Euclidean space $(\mathbb{R}^D, +, \cdot, \langle\cdot,\cdot\rangle_E)$, that is $\langle \boldsymbol{v}, \boldsymbol{u}\rangle_E := \sum_{i=1}^{D} v_i u_i$ for any two elements $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^D$, and is the core ingredient for abstracting and generalizing the concept of an angle between any two vectors and the concept of length of a vector to more general spaces. For real Hilbert spaces we can define a norm and a metric analogously, see [5]:

Definition 2. *The induced norm $\|\cdot\|_{\mathcal{H}} : \mathcal{H} \to [0, \infty)$ is defined pointwise as*

$$\|\boldsymbol{u}\|_{\mathcal{H}} := \sqrt{\langle \boldsymbol{u}, \boldsymbol{u}\rangle_{\mathcal{H}}}$$

whereas the distance $\mathrm{d} : \mathcal{H} \times \mathcal{H} \to [0, \infty)$, also called metric, between two elements $\boldsymbol{u}$ and $\boldsymbol{v}$ is defined as

$$\mathrm{d}(\boldsymbol{u}, \boldsymbol{v}) := \|\boldsymbol{u} - \boldsymbol{v}\|_{\mathcal{H}}.$$

For any two elements $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{H}$ and a scalar $\alpha \in \mathbb{R}$ the norm $\|\cdot\|_{\mathcal{H}}$ has the properties

$$\|\boldsymbol{u}\|_{\mathcal{H}} \geq 0 \text{ with equality if and only if } \boldsymbol{u} = \boldsymbol{0} \tag{2.1}$$
$$\|\alpha \boldsymbol{u}\|_{\mathcal{H}} = |\alpha| \|\boldsymbol{u}\|_{\mathcal{H}} \tag{2.2}$$
$$\|\boldsymbol{u} + \boldsymbol{v}\|_{\mathcal{H}} \leq \|\boldsymbol{u}\|_{\mathcal{H}} + \|\boldsymbol{v}\|_{\mathcal{H}}. \tag{2.3}$$

The metric $\mathrm{d}(\cdot, \cdot)$ fulfills for any three elements $\boldsymbol{u}, \boldsymbol{v}$ and $\boldsymbol{w}$ in $\mathcal{H}$:

$$\mathrm{d}(\boldsymbol{u}, \boldsymbol{v}) = 0 \iff \boldsymbol{u} = \boldsymbol{v} \tag{2.4}$$
$$\mathrm{d}(\boldsymbol{u}, \boldsymbol{v}) = \mathrm{d}(\boldsymbol{v}, \boldsymbol{u}) \tag{2.5}$$
$$\mathrm{d}(\boldsymbol{u}, \boldsymbol{v}) \leq \mathrm{d}(\boldsymbol{u}, \boldsymbol{w}) + \mathrm{d}(\boldsymbol{w}, \boldsymbol{v}). \tag{2.6}$$

An inner product leads to a very rich structure with several important consequences, see [11], of which we list only the most important in the next proposition.

Proposition 1. *For a finite dimensional real Hilbert space $(\mathcal{H}, +, \cdot, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ we have the following properties:*

- *Let D be the dimension of $\mathcal{H}$, then there exists a set of elements $\{\boldsymbol{w}_i | i = 1, \ldots, D\}$ with $\langle \boldsymbol{w}_i, \boldsymbol{w}_j \rangle_{\mathcal{H}} = 0$ and $\|\boldsymbol{w}_i\|_{\mathcal{H}} = 1$, for all $i \neq j$ - called an orthonormal basis (ONB) - such that any $\boldsymbol{u} \in \mathcal{H}$ can be written as*

$$\boldsymbol{u} = \sum_{i=1}^{D} \langle \boldsymbol{w}_i, \boldsymbol{u} \rangle_{\mathcal{H}} \boldsymbol{w}_i. \tag{2.7}$$

  .

- *For any ONB we can write $\boldsymbol{u}$ as in (2.7).*
- *For any two elements $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{H}$ the Cauchy-Schwarz inequality, $|\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\mathcal{H}}| \leq \|\boldsymbol{u}\|_{\mathcal{H}} \|\boldsymbol{v}\|_{\mathcal{H}}$, holds.*
- *If two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are orthogonal, $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\mathcal{H}} = 0$, then the Pythagorean theorem, $\|\boldsymbol{u} + \boldsymbol{v}\|^2 = \|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2$, holds.*
- *There exists a Parallelogram Law, i.e for any two elements $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{H}$:*

$$\|\boldsymbol{u} + \boldsymbol{v}\|^2 + \|\boldsymbol{u} - \boldsymbol{v}\|^2 = 2\|\boldsymbol{u}\|^2 + 2\|\boldsymbol{v}\|^2$$

- *The polarization identity holds for any $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{H}$:*

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\mathcal{H}} = \frac{1}{4} (\|\boldsymbol{u} + \boldsymbol{v}\|^2 - \|\boldsymbol{u} - \boldsymbol{v}\|^2)$$

All of the definitions and results mentioned so far in this section, with exception of the statement belonging to (2.7), which has to be replaced with an infinite analogue, remain the same when allowing for $\mathcal{H}$ to become an infinite, but separable, vector space. For more on Hilbert spaces we refer to [11].

2.3. Aitchison geometry

In this section we introduce the Aitchison geometry which is of central importance in CoDA. First, we define the set of reals we are interested in, as mentioned in Section 2.1.

Definition 3. *The D-part simplex $\mathcal{S}^D$ is defined as the set of strictly positive multivariate reals summing up to one, i.e.:*

$$\mathcal{S}^D := \left\{ (x_1, \dots, x_D)' \in \mathbb{R}_+^D \,\middle|\, \sum_{j=1}^{D} x_j = 1 \right\}. \tag{2.8}$$

An element $\boldsymbol{x}$ in $\mathcal{S}^D$ can be thought of as a vector of percentages, as described in Section 2.1, that must naturally sum up to one. Equivalently, we can also think of an $\boldsymbol{x}$ describing a discrete probability distribution with each entry being the probability of an event associated with this entry. The condition in (2.8) that all entries must sum up to one is a simple convention. Any other number than one could be taken, as will become clearer in the following. To map any point in $\mathbb{R}_+^D$ to the D-part simplex we define the closure as in [4].

Definition 4. *The closure is a function $C : \mathbb{R}_+^D \to \mathcal{S}^D$ defined pointwise by*

$$C(\boldsymbol{x}) := \frac{1}{\sum_{i=1}^{D} x_i} \boldsymbol{x}$$

for any $\boldsymbol{x} \in \mathbb{R}_+^D$.

The closure projects any multivariate point $\boldsymbol{x}$ with strictly positive entries lying on a line that starts from zero and goes through $\boldsymbol{x}$ onto the D-part simplex; as mentioned in Section 2.1 and depicted in Figure 2.1. To turn the D-part simplex into a vector space two operations, similar to addition and a scalar multiplication in $\mathbb{R}^D$, are defined.

Definition 5. *For two elements $\boldsymbol{x} = (x_1, \dots, x_D)'$ and $\boldsymbol{y} = (y_1, \dots, y_D)'$ in $\mathcal{S}^D$ as well as a scalar $\alpha \in \mathbb{R}$, we define perturbation as well as powering by*

$$\textit{Perturbation:}\quad \boldsymbol{x} \oplus \boldsymbol{y} := C((x_1 y_1, \dots, x_D y_D)')$$

$$\textit{Powering:}\quad \alpha \odot \boldsymbol{x} := C((x_1^\alpha, \dots, x_D^\alpha)').$$

The inverse of an element $\boldsymbol{x} \in \mathcal{S}^D$ is defined as $\boldsymbol{x}^{-1} := (-1) \odot \boldsymbol{x}$ and subtraction, as the inverse operation to perturbation, is defined as $\boldsymbol{x} \ominus \boldsymbol{y} := \boldsymbol{x} \oplus \boldsymbol{y}^{-1}$. The unique element that fulfills for any $\boldsymbol{x} \in \mathcal{S}^D$, $\boldsymbol{x} \oplus \boldsymbol{n} = \boldsymbol{x}$, also called the neutral element $\boldsymbol{n}$ with respect to perturbation, is given by $\boldsymbol{n} := \frac{1}{D}(1, \dots, 1)'$.

It can be shown that with these operations $(\mathcal{S}^D, \oplus, \odot)$ is a finite dimensional vector space, see [47]. The vector space structure can be further extended to a Hilbert space. The standard inner product on $\mathcal{S}^D$ is the so called Aitchison inner product and, as mentioned in the previous section, induces a norm as well as a metric on $\mathcal{S}^D$.

Definition 6. *For any two points $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{S}^D$ the Aitchison inner product is defined as*

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{A}} := \frac{1}{2D} \sum_{i,j=1}^{D} \log\left(\frac{x_i}{x_j}\right) \log\left(\frac{y_i}{y_j}\right). \tag{2.9}$$

The induced Aitchison norm of a point $\boldsymbol{x}$ *is defined as*

$$\|\boldsymbol{x}\|_{\mathcal{A}} := \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathcal{A}}}. \tag{2.10}$$

The induced distance between any two points $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^D$ *is given as*

$$\mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) := \|\boldsymbol{x} \ominus \boldsymbol{y}\|_{\mathcal{A}}. \tag{2.11}$$

The Aitchison inner product as defined in (2.9) satisfies two important properties:

- Scale invariance: One of the key properties of the Aitchison inner product (2.9) is that it is scale invariant, that is for any pair $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^D$ and a positive scalar $\lambda > 0$ we have $\langle \lambda\boldsymbol{x}, \lambda\boldsymbol{y} \rangle_{\mathcal{A}} = \langle \boldsymbol{x}, \lambda\boldsymbol{y} \rangle_{\mathcal{A}} = \langle \lambda\boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{A}} = \langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{A}}$. As discussed in Section 2.1, for percentages, a point $\boldsymbol{x} \in \mathbb{R}^D_+$ and its multiple $\lambda\boldsymbol{x}$ are considered to be the same point. Under the Aitchison geometry such two points are indistinguishable. In fact, the closure operation C could also be defined such that the entries of $C(\boldsymbol{x})$ sum up to a different constant without effecting (2.9).
- Subcompositional coherence: Subcompositional coherence can be summed up by two subpoints:
  - Any arbitrary subscomposition, that is a subvector, denoted $\boldsymbol{x}_{\text{sub}} \in \mathcal{S}^{\tilde{D}}$ with $\tilde{D} < D$, of $\boldsymbol{x} \in \mathcal{S}^D$, has smaller norm than the full vector, i.e $\|\boldsymbol{x}_{\text{sub}}\|_{\mathcal{A}} \leq \|\boldsymbol{x}\|_{\mathcal{A}}$.
  - Any subcomposition is also scale invariant.

The Aitchison inner product satisfies the usual conditions, see Section 2.2, which we give here again in the corresponding operations for a better overview.

Proposition 2. *For any three points* $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ *in* $\mathcal{S}^D$ *and* $\alpha, \beta \in \mathbb{R}$ *we have with perturbation and powering as addition and scalar multiplication that the Aitchison inner product fulfills:*

$$\langle \boldsymbol{x}, \alpha \odot \boldsymbol{y} \oplus \beta \odot \boldsymbol{z} \rangle_{\mathcal{A}} = \alpha \langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{A}} + \beta \langle \boldsymbol{x}, \boldsymbol{z} \rangle_{\mathcal{A}} \tag{2.12}$$

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{A}} = \langle \boldsymbol{y}, \boldsymbol{x} \rangle_{\mathcal{A}} \tag{2.13}$$

$$\langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathcal{A}} \geq 0 \text{ and equality if and only if } \boldsymbol{x} = \frac{1}{D}(1, \ldots, 1)'. \tag{2.14}$$

All properties as mentioned in Section 2.2 hold as the Aitchison space is a finite dimensional Hilbert space.

2.4. The quotient space view

As mentioned in the previous section, the sum condition in the definition of the D-part simplex $\mathcal{S}^D$ is somewhat arbitrary and will not affect the geometry as the Aitchison inner product (2.9) is invariant under rescaling. More precisely, a relation, see Section 2.1, can be defined on $\mathbb{R}^D_+$ by

$$\boldsymbol{x} \sim \boldsymbol{y} \iff \exists \lambda > 0 \text{ such that } \boldsymbol{y} = \lambda\boldsymbol{x},$$

which is equivalent to the relation of the log-transformed quantities

$$\log(\boldsymbol{x}) \sim \log(\boldsymbol{y}) \iff \exists \lambda > 0 \text{ such that } \log(\boldsymbol{y}) = \log(\lambda)\mathbf{1} + \log(\boldsymbol{x}), \tag{2.15}$$

where $\mathbf{1}$ is a vector of ones. We continue to use log-transformed quantities as the notation is easier. Both is equivalent however as the coordinate-wise applied logarithm is a one-to-one map from $\mathbb{R}_+^D$ to $\mathbb{R}^D$. The relation (2.15) has the following properties:

- For any $\boldsymbol{x} \in \mathbb{R}_+^D$ we have $\log(\boldsymbol{x}) \sim \log(\boldsymbol{x})$.
- If for two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}_+^D$ the relation $\log(\boldsymbol{x}) \sim \log(\boldsymbol{y})$ holds, it follows that also $\log(\boldsymbol{y}) \sim \log(\boldsymbol{x})$ holds.
- If $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}_+^D$ are such that $\log(\boldsymbol{x}) \sim \log(\boldsymbol{y})$ and $\log(\boldsymbol{y}) \sim \log(\boldsymbol{z})$ holds then this implies $\log(\boldsymbol{x}) \sim \log(\boldsymbol{z})$.

A relation that fulfills the prior three points is called an equivalence relation. An equivalence relation leads to a natural decomposition of the space it is defined on. In our case $\log\left(\mathbb{R}_+^D\right)$ can be decomposed into a union of non-overlapping subsets defined by $[\log(\boldsymbol{x})] := \{\log(\boldsymbol{y}) | \log(\boldsymbol{y}) = \log(\lambda)\mathbf{1} + \log(\boldsymbol{x}), \lambda > 0\}$ for any $\boldsymbol{x} \in \mathbb{R}_+^D$. The sets $[\log(\boldsymbol{x})]$ are the log-transformed rays as mentioned in Section 2.1. We can fix for each such ray an arbitrary representative without losing any information as for two elements $\log(\boldsymbol{x}), \log(\boldsymbol{y})$ from the same equivalence class we get $[\log(\boldsymbol{x})] = [\log(\boldsymbol{y})]$. The whole set of equivalence classes $\{[\log(\boldsymbol{x})] | \boldsymbol{x} \in \mathbb{R}_+^D\}$ can then be described by these representatives. To make the connections to $\mathcal{S}^D$ we could pick for each class a representative $\boldsymbol{x} \in \mathbb{R}_+^D$ with $\sum_{i=1}^D x_i = 1$, such that $\{[\log(\boldsymbol{x})] | \boldsymbol{x} \in \mathbb{R}_+^D\} = \{[\log(\boldsymbol{x})] | \sum_{i=1}^D x_i = 1\}$. This is possible as for any $\boldsymbol{x} \in \mathbb{R}_+^D$ we have $\log(\boldsymbol{x}) = \log\left(\sum_{i=1}^D x_i\right)\mathbf{1} + \log\left(\frac{1}{\sum_{i=1}^D x_i}\boldsymbol{x}\right)$ and therefore $\log(\boldsymbol{x})$ and $\log(C(\boldsymbol{x}))$ belong to the same equivalence class. Therefore, after taking the exponential, we can see $\mathcal{S}^D$ as the space of $\{[\log(\boldsymbol{x})] | x \in \mathbb{R}_+^D\}$ restricted to special representatives. In the theory of quotient spaces, see [52], one can define scalar multiplication and addition on the equivalence classes, such that these operations are invariant to the chosen representatives. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}_+^D$ and $\alpha \in \mathbb{R}$, one defines

$$[\log(\boldsymbol{x})] + [\log(\boldsymbol{y})] := [\log(\boldsymbol{x}) + \log(\boldsymbol{y})] \text{ and } \alpha[\log(\boldsymbol{x})] := [\alpha \log(\boldsymbol{x})]. \tag{2.16}$$

As mentioned before the elements $\log(\boldsymbol{x})$ and $\log(C(\boldsymbol{x}))$ are in the same equivalence class and so we can write $[\log(\boldsymbol{x})]$ equivalently as $[\log(C(\boldsymbol{x}))]$. The right-hand side of (2.16) respectively reads thus as $[\log(\boldsymbol{x}) + \log(\boldsymbol{y})] = [\log(\boldsymbol{x}\boldsymbol{y})] = [\log(C(\boldsymbol{x}\boldsymbol{y}))]$ and $[\alpha \log(\boldsymbol{x})] = [\log(\boldsymbol{x}^\alpha)] = [\log(C(\boldsymbol{x}^\alpha))]$. Taking the exponential we can see that this is precisely the definition of perturbation and powering with a specific representative. Such a quotient space approach was considered in [7]. Further constructions, such as the Aitchison inner product and the Aitchison norm can also be recovered in this setting. For example, to motivate the norm one can look at the fact that the log-differences $\log(\boldsymbol{x}) - \log(\boldsymbol{y})$ of any two elements, $\boldsymbol{x}$ and $\boldsymbol{y}$, which are in the same equivalence class, are an element of the space spanned by $\mathbf{1}$. As on this space any two elements are regarded as being equal it seems natural to measure angle, distance and length of an element in the direction orthogonal to this vector. Projecting any element $\log(\boldsymbol{x})$ onto the orthogonal of this space leads to taking the (Euclidean) length of $\log(\boldsymbol{x}) - \frac{1}{D}\sum_{i=1}^D \log(x_i)\mathbf{1}$; this is precisely the Aitchision norm of $\boldsymbol{x}$.

In light of this quotient approach we can see that the condition in $\mathcal{S}^D$ is arbitrary and can be discarded in principle, i.e. any $\boldsymbol{x} \in \mathbb{R}_+^D$ that is rescaled by a constant is considered to be the same element when using the Aitchison norm respectively inner product.

2.5. Important mappings

A number of interpretable linear transformations from $\mathcal{S}^D$ to the multivariate reals exist. The main ingredient of the Aitchison inner product (2.9) are the $D(D-1)$ pairwise log-ratios. Therefore, maps which are scale invariant and linear and whose coordinates can recover all log-ratios, when linearly combined, carry all the information necessary. Historically, one of the oldest such transformations is the additive log-ratio map (alr_k) indexed by $k \in \{1,\dots,D\}$.

Definition 7. *For a fixed index $k \in \{1,\dots,D\}$, the* alr_k *map is defined from $\mathcal{S}^D$ to $\mathbb{R}^{D-1}$ coordinate-wise by*

$$\mathrm{alr}_k(\boldsymbol{x}) := \left(\log\left(\frac{x_1}{x_k}\right),\dots,\log\left(\frac{x_{k-1}}{x_k}\right),\dots,\log\left(\frac{x_{k+1}}{x_k}\right),\dots,\log\left(\frac{x_D}{x_k}\right)\right)' \in \mathbb{R}^{D-1}. \tag{2.17}$$

It is easy to see that any log-ratio $\log\left(\frac{x_i}{x_j}\right)$ can be written as a combination of the coordinates of $\mathrm{alr}_k(\boldsymbol{x})$, $\log\left(\frac{x_i}{x_j}\right) = \log\left(\frac{x_i}{x_k}\right) - \log\left(\frac{x_j}{x_k}\right)$, and contains therefore all important information. One might be thus lead to think that usual statistical methods in $\mathbb{R}^{D-1}$ could be applied. However, one can show that alr_k is, for general $\boldsymbol{x},\boldsymbol{y} \in \mathcal{S}^D$, not distance nor angle preserving, i.e. $\langle \boldsymbol{x},\boldsymbol{y}\rangle_{\mathcal{A}} \neq \langle \mathrm{alr}_k(\boldsymbol{x}), \mathrm{alr}_k(\boldsymbol{y})\rangle_E$, and should therefore, in cases that a method is based on the Euclidean geometry, not be used. Due to this weakness and the additional disadvantage of having to choose a fixed index k the alr_k map is not often utilized. Instead far more appropriate maps have been investigated. An example is the centered log-ratio map (clr) which does preserve distances and angles of $\boldsymbol{x},\boldsymbol{y} \in \mathcal{S}^D$ with respect to the Euclidean geometry.

Definition 8. *The* clr *map is defined from $\mathcal{S}^D$ to $\mathbb{R}^D$ coordinate-wise by*

$$\mathrm{clr}(\boldsymbol{x}) := \left(\log\left(\frac{x_1}{\sqrt[D]{\prod_{j=1}^D x_j}}\right),\dots,\log\left(\frac{x_D}{\sqrt[D]{\prod_{j=1}^D x_j}}\right)\right)'. \tag{2.18}$$

Again, any log-ratio can be written in terms of coordinates of the clr map, $\log\left(\frac{x_i}{x_j}\right) = \mathrm{clr}(\boldsymbol{x})_i - \mathrm{clr}(\boldsymbol{x})_j$, which means that the clr map carries all the relevant information. Each coordinate of the latter can be written as a sum of log-ratios,

$$\mathrm{clr}(\boldsymbol{x})_i = \log\left(\frac{x_i}{\sqrt[D]{\prod_{j=1}^D x_j}}\right) = \frac{1}{D}\left(\log\frac{x_i}{x_1} + \dots + \log\frac{x_i}{x_{i-1}} + \log\frac{x_i}{x_{i+1}} + \dots + \log\frac{x_i}{x_D}\right),$$

and so the i-th coordinate can be interpreted as an average of all parts involved with said coordinate. More compactly, we can write (2.18) also as

$$\mathrm{clr}(\boldsymbol{x}) = \mathbf{L}\log(\boldsymbol{x}), \tag{2.19}$$

with the centering matrix $\mathbf{L} := \mathbf{I} - \frac{1}{D}\mathbf{1}\mathbf{1}'$, see [52], having the property $\mathbf{L}^2 = \mathbf{L}$. It can be shown that the clr map respects the perturbation and powering operation

$$\mathrm{clr}(\boldsymbol{x} \oplus \boldsymbol{y}) = \mathrm{clr}(\boldsymbol{x}) + \mathrm{clr}(\boldsymbol{y}),\ \mathrm{clr}(\alpha \odot \boldsymbol{x}) = \alpha\,\mathrm{clr}(\boldsymbol{x}), \tag{2.20}$$

for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^D$ and $\alpha \in \mathbb{R}$, and is angle and length preserving with respect to the Euclidean inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{A}} = \langle \operatorname{clr}(\boldsymbol{x}), \operatorname{clr}(\boldsymbol{y}) \rangle_E. \tag{2.21}$$

As the D-part simplex $\mathcal{S}^D$ is a subset of a D-dimensional space, $\exp(\mathbb{R}^D) = \mathbb{R}^D_+$, with a sum constraint, it can be shown that this reduces the dimension to $D-1$. The clr map however maps $\mathcal{S}^D$ into $\mathbb{R}^D$ and not $\mathbb{R}^{D-1}$ and can therefore not be bijective. To construct a bijective map it helps to notice that the sum of all $\operatorname{clr}(\boldsymbol{x})$ coordinates is zero, $\langle \operatorname{clr}(\boldsymbol{x}), \mathbf{1} \rangle_E = 0$, for any $\boldsymbol{x} \in \mathcal{S}^D$. As the clr map contains all the information about all the log-ratios it spans an at least $D-1$ dimensional space and so it can be deduced that the image of the clr map spans the whole of $\{\boldsymbol{z} \in \mathbb{R}^{D-1} | \langle \boldsymbol{z}, \mathbf{1} \rangle_E = 0\}$. To construct a bijective map the idea is therefore to take an orthogonal basis of $\{\boldsymbol{z} \in \mathbb{R}^{D-1} | \langle \boldsymbol{z}, \mathbf{1} \rangle_E = 0\}$, typically denoted by vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{D-1}$ and collected column-wise into the matrix $\mathbf{V} \in \mathbb{R}^{D \times (D-1)}$, and express for any $\boldsymbol{x} \in \mathcal{S}^D$ the vector $\operatorname{clr}(\boldsymbol{x})$ as a linear combination of the columns of $\mathbf{V}$. This construction leads to the isometric log-ratio map ($\operatorname{ilr}_{\mathbf{V}}$) which can be defined as the unique vector solving $\operatorname{clr}(\boldsymbol{x}) = \mathbf{V} \operatorname{ilr}_{\mathbf{V}}(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{S}^D$:

Definition 9. *For a given matrix $\mathbf{V} \in \mathbb{R}^{D \times (D-1)}$ with columns being an orthogonal basis of $\{\boldsymbol{z} \in \mathbb{R}^{D-1} | \langle \boldsymbol{z}, \mathbf{1} \rangle_E = 0\}$, we define the $\operatorname{ilr}_{\mathbf{V}}$ map from $\mathcal{S}^D$ to $\mathbb{R}^{D-1}$ pointwise as the unique solution to*

$$\operatorname{clr}(\boldsymbol{x}) = \mathbf{V} \operatorname{ilr}_{\mathbf{V}}(\boldsymbol{x}). \tag{2.22}$$

From the definition of the $\operatorname{ilr}_{\mathbf{V}}$ map and the fact that $\mathbf{V}$ is an orthogonal matrix, $\mathbf{V}'\mathbf{V} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix, we can deduce that $\operatorname{ilr}_{\mathbf{V}}(\boldsymbol{x}) = \mathbf{V}' \operatorname{clr}(\boldsymbol{x})$ holds for any $\boldsymbol{x} \in \mathcal{S}^D$. Furthermore, we can also directly conclude

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{A}} = \langle \operatorname{clr}(\boldsymbol{x}), \operatorname{clr}(\boldsymbol{y}) \rangle_E = \langle \mathbf{V} \operatorname{ilr}_{\mathbf{V}}(\boldsymbol{x}), \mathbf{V} \operatorname{ilr}_{\mathbf{V}}(\boldsymbol{y}) \rangle_E = \langle \operatorname{ilr}_{\mathbf{V}}(\boldsymbol{x}), \operatorname{ilr}_{\mathbf{V}}(\boldsymbol{y}) \rangle_E, \tag{2.23}$$

and so the $\operatorname{ilr}_{\mathbf{V}}$ map is also distance and angle preserving. By definition properties (2.20) will still hold. All in all, we have the following proposition.

Proposition 3. *For any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^D$ and $\alpha \in \mathbb{R}$ and fixed matrices $\mathbf{V}$ and $\tilde{\mathbf{V}}$ with orthogonal columns spanning $\{\boldsymbol{z} \in \mathbb{R}^{D-1} | \langle \boldsymbol{z}, \mathbf{1} \rangle_E = 0\}$ we have:*

- $\operatorname{ilr}_{\mathbf{V}}(\boldsymbol{x}) = \mathbf{V}' \operatorname{clr}(\boldsymbol{x})$ *and* $\operatorname{clr}(\boldsymbol{x}) = \mathbf{V} \operatorname{ilr}_{\mathbf{V}}(\boldsymbol{x})$.
- *A basis change in $\{\boldsymbol{z} \in \mathbb{R}^{D-1} | \langle \boldsymbol{z}, \mathbf{1} \rangle_E = 0\}$ changes the* ilr *map correspondingly*

$$\operatorname{ilr}_{\tilde{\mathbf{V}}}(\boldsymbol{x}) = \tilde{\mathbf{V}}' \mathbf{V} \operatorname{ilr}_{\mathbf{V}}(\boldsymbol{x}). \tag{2.24}$$

- *Perturbation and powering are mapped accordingly*

$$\operatorname{ilr}_{\mathbf{V}}(\boldsymbol{x} \oplus \boldsymbol{y}) = \operatorname{ilr}_{\mathbf{V}}(\boldsymbol{x}) + \operatorname{ilr}_{\mathbf{V}}(\boldsymbol{y}) \text{ and } \operatorname{ilr}_{\mathbf{V}}(\alpha \odot \boldsymbol{x}) = \alpha \operatorname{ilr}_{\mathbf{V}}(\boldsymbol{x}). \tag{2.25}$$

- *Angles and distances are preserved*

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathcal{A}} = \langle \operatorname{ilr}_{\mathbf{V}}(\boldsymbol{x}), \operatorname{ilr}_{\mathbf{V}}(\boldsymbol{y}) \rangle_E. \tag{2.26}$$

- *The inverse of* $\mathrm{ilr}_{\mathbf{V}}$ *maps from* $\mathbb{R}^{D-1}$ *to* $\mathcal{S}^D$ *and is defined pointwise for any fixed* $z \in \mathbb{R}^{D-1}$ *by*

$$\mathrm{ilr}_{\mathbf{V}}^{-1}(z) = C(\exp(\mathbf{V}z)).$$

Many choices of orthogonal basis vectors collected in $\mathbf{V}$ exist. If the emphasis is on interpretability, special bases are sought after. The most important and commonly used ones are known as pivot coordinates, symmetric pivot coordinates and balances, see [19].

2.6. Compositional regression

In the following we will write $\bigoplus_{k=1}^{K} \boldsymbol{x}_k$ for the perturbation over K compositional vectors $\boldsymbol{x}_k \in \mathcal{S}^D$, $k = 1, \ldots, K$. Many statistical concepts can be phrased in the Hilbert space framework. As the Aitchison space is a Hilbert space, many concepts can directly be transferred as such. In this section we give a quick review of compositional regression models. Three different types of commonly used models come to mind, although more settings exist:

- Regressing a real variable $y \in \mathbb{R}$ onto a compositional vector $\boldsymbol{x} \in \mathcal{S}^D$.
- Regressing a compositional variable $\boldsymbol{y} \in \mathcal{S}^{\tilde{D}}$ onto a real vector $\boldsymbol{x} \in \mathbb{R}^D$.
- Regressing a compositional variable $\boldsymbol{y} \in \mathcal{S}^{\tilde{D}}$ onto a compositional vector $\boldsymbol{x} \in \mathcal{S}^D$.

Before looking at each case separately it is informative to look at linear models in Hilbert spaces. Any linear function $\mathbf{F}$ from one Hilbert space $\mathcal{H}_1$, with dimension D, to another Hilbert space $\mathcal{H}_2$, with dimension $\tilde{D}$, can be written in different forms as

$$\mathbf{F}(\boldsymbol{x}) = \sum_{l=1}^{D} \mathbf{F}(\boldsymbol{u}_l)\langle \boldsymbol{u}_l, \boldsymbol{x}\rangle_{\mathcal{H}_1} = \sum_{l=1}^{D}\sum_{m=1}^{\tilde{D}} \langle \tilde{\boldsymbol{u}}_m, \mathbf{F}(\boldsymbol{u}_l)\rangle_{\mathcal{H}_2}\langle \boldsymbol{u}_l, \boldsymbol{x}\rangle_{\mathcal{H}_1}\tilde{\boldsymbol{u}}_m \tag{2.27}$$

$$= \sum_{l=1}^{D}\sum_{m=1}^{\tilde{D}} F_{ml}\langle \boldsymbol{u}_l, \boldsymbol{x}\rangle_{\mathcal{H}_1}\tilde{\boldsymbol{u}}_m \tag{2.28}$$

$$= \sum_{l=1}^{D} \langle \boldsymbol{u}_l, \boldsymbol{x}\rangle_{\mathcal{H}_1}\hat{\boldsymbol{b}}_l \tag{2.29}$$

$$= \sum_{m=1}^{\tilde{D}} \langle \boldsymbol{b}_m, \boldsymbol{x}\rangle_{\mathcal{H}_1}\tilde{\boldsymbol{u}}_m, \tag{2.30}$$

where the sums are understood in the respective Hilbert spaces, the vectors $\boldsymbol{u}_l$ are an orthogonal basis in $\mathcal{H}_1$, the vectors $\tilde{\boldsymbol{u}}_m$ are an orthogonal basis in $\mathcal{H}_2$, $F_{ml} := \langle \tilde{\boldsymbol{u}}_m, \mathbf{F}(\boldsymbol{u}_l)\rangle_{\mathcal{H}_2}$ and $\boldsymbol{b}_m := \sum_{l=1}^{D} F_{ml}\boldsymbol{u}_l$ as well as $\hat{\boldsymbol{b}}_l := \sum_{m=1}^{\tilde{D}} F_{ml}\tilde{\boldsymbol{u}}_m$ are coefficient vectors. This chain of equality is useful as any linear function can be written in the forms of (2.27) - (2.30) where, in a regression context, we need to find the coefficient vectors $\boldsymbol{b}_m$ or $\hat{\boldsymbol{b}}_l$ after having fixed a basis. From this we can directly deduce the three compositional regression cases as mentioned above:

- Using (2.30) and the fact that the basis of $\mathcal{H}_2 = \mathbb{R}$ is the scalar one, we can see that any linear function from $\mathcal{S}^D$ to $\mathbb{R}$ is given by $\langle \boldsymbol{a}, \boldsymbol{x} \rangle_{\mathcal{A}}$. For observed samples $(y_n, \boldsymbol{x}_n) \in \mathbb{R} \times \mathcal{S}^D$, $n = 1, \ldots, N$, and errors ϵ_n, the linear regression model takes the form

$$y_n = \mu + \langle \boldsymbol{a}, \boldsymbol{x}_n \rangle_{\mathcal{A}} + \epsilon_n, \tag{2.31}$$

where $\mu \in \mathbb{R}$ and $\boldsymbol{a} \in \mathcal{S}^D$ are the coefficients to be estimated. To solve (2.31) one uses property (2.26) to reformulate the model equation as $y_n = \mu + \langle \mathrm{ilr}_{\mathbf{V}}(\boldsymbol{a}), \mathrm{ilr}_{\mathbf{V}}(\boldsymbol{x}_n) \rangle_E + \epsilon_n$ which can then be solved by standard methods.

- Using (2.29) and as orthonormal basis in $\mathcal{H}_1 = \mathbb{R}^D$ the standard canonical basis $\boldsymbol{e}_l$, which are vectors of zeros with one in the l-th entry - so that $\langle \boldsymbol{e}_l, \boldsymbol{x} \rangle_E = x_l$ - we can see that any linear function $\mathbf{F}$ from $\mathbb{R}^D$ to $\mathcal{S}^{\tilde{D}}$ can be written as

$$\mathbf{F}(\boldsymbol{x}) = \bigoplus_{l=1}^{D} x_l \odot \hat{\boldsymbol{b}}_l.$$

For observed samples $(\boldsymbol{y}_n, \boldsymbol{x}_n) \in \mathcal{S}^{\tilde{D}} \times \mathbb{R}^D$, $n = 1, \ldots, N$, and errors $\boldsymbol{\epsilon}_n$, the linear regression model takes the form

$$\boldsymbol{y}_n = \boldsymbol{\mu} \oplus \bigoplus_{l=1}^{D} x_{nl} \odot \hat{\boldsymbol{b}}_l \oplus \boldsymbol{\epsilon}_n \tag{2.32}$$

where $\boldsymbol{\mu} \in \mathcal{S}^{\tilde{D}}$ and $\hat{\boldsymbol{b}}_l \in \mathcal{S}^{\tilde{D}}$ are coefficient vectors to be estimated. To solve (2.32) we can use (2.25) reformulating the former as

$$\mathrm{ilr}_{\mathbf{V}}(\boldsymbol{y}_n) = \mathrm{ilr}_{\mathbf{V}}(\boldsymbol{\mu}) + \sum_{l=1}^{D} \mathrm{ilr}_{\mathbf{V}}(\hat{\boldsymbol{b}}_l) x_{nl} + \mathrm{ilr}_{\mathbf{V}}(\boldsymbol{\epsilon}_n),$$

which is a standard multivariate D dimensional regression model with multivariate $\tilde{D} - 1$ dimensional real response that can easily be solved by standard methods.

- The representation in (2.30) can be used again to deduce that any linear function $\mathbf{F}$ from $\mathcal{S}^D$ to $\mathcal{S}^{\tilde{D}}$ can be written as

$$\mathbf{F}(\boldsymbol{x}) = C(\exp(\mathbf{A} \log(\boldsymbol{x}))) \tag{2.33}$$

with a matrix $\mathbf{A} \in \mathbb{R}^{\tilde{D} \times D}$ such that $\mathbf{A1} = \mathbf{0}$ and $\mathbf{1}'\mathbf{A} = \mathbf{0}'$ holds; to show this is however more cumbersome and will not be done here, instead we refer to [18]. For observed samples $(\boldsymbol{y}_n, \boldsymbol{x}_n) \in \mathcal{S}^{\tilde{D}} \times \mathcal{S}^D$, $n = 1, \ldots, N$, and errors $\boldsymbol{\epsilon}_n$, the linear regression model takes the form

$$\boldsymbol{y}_n = \boldsymbol{\mu} \oplus \exp(\mathbf{A} \log(\boldsymbol{x}_n)) \oplus \boldsymbol{\epsilon}_n,$$

where $\boldsymbol{\mu} \in \mathcal{S}^{\tilde{D}}$ and $\mathbf{A} \in \mathbb{R}^{\tilde{D} \times D}$, with $\mathbf{A1} = \mathbf{0}$ and $\mathbf{1}'\mathbf{A} = \mathbf{0}$, are the coefficients to be estimated. The latter can be phrased again in ilr coordinates and leads to $\mathrm{ilr}_{\mathbf{V}_2}(\boldsymbol{y}_n) = \mathrm{ilr}_{\mathbf{V}_2}(\boldsymbol{\mu}) + \mathbf{V}_2'\mathbf{A}\mathbf{V}_1 \log(\boldsymbol{x}_n) + \mathrm{ilr}_{\mathbf{V}_2}(\boldsymbol{\epsilon}_n)$ where $\mathbf{V}_1$ corresponds to the ilr taken in $\mathcal{S}^D$ and $\mathbf{V}_2$ respectively to the ilr in $\mathcal{S}^{\tilde{D}}$, see [8]. This can be solved by standard methods after setting $\mathbf{B} := \mathbf{V}_2'\mathbf{A}\mathbf{V}_1$, where $\mathbf{B}$ can vary without constraints. For a more thorough introduction to such models and extensions we refer to [8].

Any of the models mentioned above display difficulties when it comes to interpreting the estimated coefficients. However, as the linear models for all three cases above do not depend on which ilr transform is used - they can be phrased solely in a Hilbert space setting - we can directly switch from one coordinate map, e.g. pivot coordinates, to another, for example balances. Inference on the estimated coefficient can also be done. One needs however to pay attention to which coordinate map is used to get meaningful results.

3. Geographically dependent log-ratios

In this chapter we sum up the main ideas and concepts of Publication I and II. In both papers we developed a new method for the detection of mineralization given the concentration of various chemical elements sampled on a linear transect (Publication I) or a whole grid (Publication II). The method is based on Generalized Additive Models (GAMs) and their corresponding curvature.

3.1. Modeling scale invariance in a geographical setting

Assume that we are in a geographical setting in which we are interested to model the dependence of a multivariate quantity $\boldsymbol{q}$ on its location $\boldsymbol{s} \in \Omega$, where Ω is either an interval in $\mathbb{R}$ or a whole area in $\mathbb{R}^2$. For each entry q_l of $\boldsymbol{q}$ we are given erroneous samples q_{nl}, $n = 1, \ldots, N$, at some location $\boldsymbol{s}_{nl} \in \Omega$. To construct a global estimator of $\boldsymbol{q}$, we regress for each entry separately, the measurements q_{nl} onto the locations $\boldsymbol{s}_{nl}$, $n = 1, \ldots, N$, using GAMs. One core assumption to be able to use GAMs is that $q_l(\boldsymbol{s})$ follows, for fixed $\boldsymbol{s} \in \Omega$, a distribution with density belonging to the exponential family

$$f(q_l(\boldsymbol{s})|\theta) := \exp\left(\frac{\theta q_l(\boldsymbol{s}) - b(\theta)}{a(\psi)} + c(q_l(\boldsymbol{s}), \psi)\right),$$

where θ is an unknown parameter, a, b, c are fixed functions and ψ is a dispersion parameter, all depending on the chosen distribution, see [45] or [13]. Under appropriate assumptions it can be shown that

$$\mathbb{E}(q_l(\boldsymbol{s})) = b'(\theta) \tag{3.1}$$

$$\mathbb{V}(q_l(\boldsymbol{s})) = b''(\theta)a(\psi), \tag{3.2}$$

where $\mathbb{E}$ denotes the expectation, $\mathbb{V}$ the variance, the $'$ superscript the derivative, and $a(\psi) = \frac{\psi}{\omega(\boldsymbol{s})}$ is taken, hold. The importance of Equation (3.1) is that if we model θ as some function dependent on $\boldsymbol{s}$ the range of the mean of $q_l(\boldsymbol{s})$ will always match the range of our model $b'(\theta(\boldsymbol{s}))$. Usually a slightly more general version is preferred, modeling

$$h(\mathbb{E}(q(\boldsymbol{s}))) = \eta(\boldsymbol{s}) \tag{3.3}$$

for a fixed smooth monotonic function h with appropriate domain and a smooth η, instead; see [64] for a more thorough explanation. Under this model one can write the log-likelihood function of $q_l(\boldsymbol{s})$ in dependence of $\eta(\boldsymbol{s})$, i.e $l(q_l(\boldsymbol{s})|\eta(\boldsymbol{s}))$, instead of $\theta(\boldsymbol{s})$. To get a smooth estimate $\hat{\eta}$ of η, which is parameterized as

$$\eta(\boldsymbol{s}) = \sum_{k=1}^{K} \beta_k m_k(\boldsymbol{s}) = \boldsymbol{\beta}' \boldsymbol{m}(\boldsymbol{s})$$

where $\boldsymbol{\beta} := (\beta_k)_{k=1,\dots,K}$ is a coefficient vector that is to be estimated and where m_k are basis functions, collected in $\boldsymbol{m} := (m_k)_{k=1,\dots,K}$, commonly corresponding to the cubic B-spline basis, see [28], Duchon splines, see [15], tensor product splines, see [65] or soap film smoothers, see [66], a penalized log-likelihood problem is solved:

$$\hat{\boldsymbol{\beta}}' := \underset{\boldsymbol{\beta}\in\mathbb{R}^K}{\arg\max} \sum_{n=1}^{N} w_n^l l(q_{nl}|\boldsymbol{\beta}'\boldsymbol{m}(\boldsymbol{s}_n)) - \lambda\mathcal{P}(\boldsymbol{\beta}). \tag{3.4}$$

In (3.4), $\lambda > 0$ is a smoothing parameter, w_n^l are weights to down or up-weigh certain samples, and $\mathcal{P}(\boldsymbol{\beta}) = \boldsymbol{\beta}'\mathbf{S}\boldsymbol{\beta}$ is a quadratic penalty, with matrix $\mathbf{S} \in \mathbb{R}^{K\times K}$ that controls the smoothness of η. In Publication I we considered the penalty $\mathcal{P}$ with corresponding matrix-entries

$$\mathbf{S}_{ij} := \int_\Omega m_i''(s)m_j''(s)ds,$$

over an interval $\Omega \in \mathbb{R}$ and in Publication II respectively the corresponding matrix-entries

$$\mathbf{S}_{ij} := \int_\Omega \left(\frac{\partial^2}{\partial s_1}m_i(\boldsymbol{s}) + \frac{\partial^2}{\partial s_2}m_i(\boldsymbol{s})\right)\left(\frac{\partial^2}{\partial s_1}m_j(\boldsymbol{s}) + \frac{\partial^2}{\partial s_2}m_j(\boldsymbol{s})\right)ds_1 ds_2$$

over an area $\Omega \in \mathbb{R}^2$, for $i, j \in \{1,\dots,K\}$. We refer to [64] for more details on different smoothing penalties and models. In both publications we assume $q_l(\boldsymbol{s})$ to follow a Tweedie distribution, see [36]. The latter is very flexible and its variance takes the form

$$\mathbb{V}(q_l(\boldsymbol{s})) = \mathbb{E}(q_l(\boldsymbol{s}))^p \frac{\psi}{w(\boldsymbol{s})}$$

for a fixed $p \in (1,2)$. Together with the log-link function, $h \equiv \log$, we obtain a scale equivariant predictor for the mean, i.e rescaling of $q_l(\boldsymbol{s})$ by a constant rescales $h^{-1}(\eta(\boldsymbol{s}))$ accordingly; this can be deduced directly from (3.3) and the fact that the Tweedie distribution is closed under rescaling for any such fixed p, see [36]. Modeling each $q_l(\boldsymbol{s})$ of a strictly positive multivariate quantity $\boldsymbol{q}(\boldsymbol{s}) \in \mathbb{R}_+^D$ in the above way leads to pairwise log-ratio functions, $\log\left(\frac{q_l(\boldsymbol{s})}{q_t(\boldsymbol{s})}\right)$ for $l, t = 1,\dots,D$, which are invariant under rescaling.

3.2. An application to geochemical exploration

In Publication I and II we looked at chemical concentrations that belonged to geomchemical measurements sampled over a linear transect (Publication I) and over a grid (Publication II). The overall goal was to construct a measure which indicates possible mineralization that typically occurs in situations where some particular chemical elements display comparatively high concentrations. As a preprocessing step the methodology of the previous section is applied to each of the D-many chemical elements separately, resulting in an estimated multivariate function $\hat{\boldsymbol{q}}$ over Ω, each coordinate $\hat{q}_l$, with $l = 1,\dots,D$, describing a smooth predictor of the mean of the l-th element. The main idea in both publications is to construct a measure based on the curvature of the pairwise log-ratio functions

$$\boldsymbol{s} \mapsto \log\left(\frac{\hat{q}_l(\boldsymbol{s})}{\hat{q}_t(\boldsymbol{s})}\right) \text{ for } \boldsymbol{s} \in \Omega, \tag{3.5}$$

for any $l, t = 1, \ldots, D$. The curvature, where in Publication I we used the standard one-dimensional curvature, see [38], and in Publication II the mean curvature, see [12], can be thought of as a measure of how quickly a signal changes in $\boldsymbol{s}$. Fast spatial changes of (3.5) in $\boldsymbol{s}$ could be an indication of interesting mineralizations and might reveal interesting pairs of elements in relation to such; this is also the reason for using GAMs in the preprocessing step to estimate $\boldsymbol{q}$ instead of Kriging, see [46], as the penalty, $\mathcal{P}$ in (3.4), leads to estimators that display, loosely speaking, large curvature where necessary. Based on the curvature we constructed a measure that gives a score in $[0, 1]$ for each log-ratio function (3.5), whereas the closer it is to one the comparatively more spatial variation a log-ratio function displays.

Publication I: Identification of mineralization in geochemistry along a transect based on the spatial curvature of log-ratios

Summary

In Publication I a new method to discover interesting chemical elements that indicate possible mineralizations, as well as their position along a transect, is developed. The method is based, in a first step, on smoothing the chemical concentrations over the transect with Generalized Additive models. The smoothed concentrations are then used to construct a measure based on the curvature of all possible log-ratio functions.

Bibliographic information

D. Mikšová, C. Rieser, and P. Filzmoser. Identification of mineralization in geochemistry along a transect based on the spatial curvature of log-ratios. Mathematical Geosciences 53, 1513-1533, 2021. https://doi.org/10.1007/s11004-021-09930-4.

Author's contribution

C. Rieser contributed to the development of the methodology, implemented most of the methods in R and wrote Chapter 2 and 3 of the corresponding paper.

Publication II: Identification of mineralization in geochemistry for grid sampling using Generalized Additive Models

Summary

Publication II is an extension of Publication I in regard to the measurements, i.e. the concentrations of chemical elements, being sampled on a whole two dimensional grid opposed to a linear transect. As in Publication I, a measure to detect interesting chemical elements related to possible mineralizations and their position is developed. This measure is based on a generalized notion of curvature appropriate for functions describing a surface. Similar to Publication I, surface functions for all possible log-ratios are constructed from a smoothed version of the individual elements.

Bibliographic information

D. Mikšová, C. Rieser, P. Filzmoser, M. Middleton, and R. Sutinen. Identification of mineralization in geochemistry for grid sampling using Generalized Additive Models. Mathematical Geosciences 53, 1861–1880, 2021. https://doi.org/10.1007/s11004-021-09929-x.

Author's contribution

C. Rieser developed the methodology of the extension in relation to Generalized Additive Models and the curvature measure based on differential geometry. Additionally, C. Rieser implemented much of the R Code and wrote the Methodology as well as the Algorithm section of the corresponding paper.

4. Detecting trends in data

In this chapter we recapture the main ideas of Publication III. In the latter we looked at compositional time series and the fitting of piecewise (compositional) linear trends to multivariate time series in $\mathcal{S}^D$. The method extends univariate l_1 trend filtering , see [37], to a compositional setting.

4.1. Univariate trend filtering

Given a univariate time series $x_t \in \mathbb{R}$ dependent on a time stamp $t = 1, \dots, T$, univariate trend filtering is a technique to fit piecewise linear trends to x_t. It is very similar to the widely used Hodrick–Prescott filter, see [29], which reads as the solution to

$$\hat{a}_t := \underset{a_t \in \mathbb{R}}{\arg\min} \sum_{t=1}^{T} (x_t - a_t)^2 + \lambda \sum_{t=3}^{T} (a_t - 2a_{t-1} + a_{t-2})^2 \tag{4.1}$$

for a $\lambda > 0$ that controls the variation of the estimate a_t. In [37] the authors consider replacing the quadratic penalty in the second sum of (4.1) by an l_1 penalty leading to the reformulated problem

$$\hat{a}_t := \underset{a_t \in \mathbb{R}}{\arg\min} \sum_{t=1}^{T} (x_t - a_t)^2 + \lambda \sum_{t=3}^{T} |a_t - 2a_{t-1} + a_{t-2}|. \tag{4.2}$$

This change leads to a shrinkage of $a_t - 2a_{t-1} + a_{t-2}$, similar as for the Lasso estimator, see [59], shrinking many of the latter to zero with growing λ. It is easy to see that when for consecutive times, $t = t_0, t_1, \dots, t_l$, with $l > 1$, $a_t - 2a_{t-1} + a_{t-2} = 0$ holds, we have that a_t is a linear function in t for these times, $a_t = at + b$, for some $a, b \in \mathbb{R}$. Figure 4.1 shows an artificial example, with simulated points x_t, for $t = 1, \dots, 101$, in black, and the trend filtering fit, $\hat{a}_t$ in red, with automatic estimate of λ as implemented in the R package [6].

4.2. Compositional trend filtering

Since its publication, a multitude of extensions to l_1 trend filtering have appeared [63], [61] or [55]. In Publication III we consider reformulating (4.2) in a compositional context. The goal is to extract piecewise linear trends of a compositional time series $\boldsymbol{x}_t \in \mathbb{S}^D$, i.e $\boldsymbol{a}_t = \boldsymbol{a} \odot t \oplus \boldsymbol{b}$, for some $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{S}^D$, and consecutive $t = t_0, t_1, \dots, t_l$, compare with Section 2.6. As the l_1 penalty in (4.2) has a shrinkage effect, we replaced it with $\|\boldsymbol{a}_t \ominus 2 \odot \boldsymbol{a}_{t-1} \oplus \boldsymbol{a}_{t-2}\|_{\mathcal{A}}$ to get the compositional multivariate pendant. All in all the problem then reads as

$$\hat{a}_t := \underset{\boldsymbol{a}_t \in \mathbb{S}^D}{\arg\min} \sum_{t=1}^{T} \|\boldsymbol{x}_t \ominus \boldsymbol{a}_t\|_{\mathcal{A}}^2 + \lambda \sum_{t=3}^{T} \|\boldsymbol{a}_t \ominus 2 \odot \boldsymbol{a}_{t-1} \oplus \boldsymbol{a}_{t-2}\|_{\mathcal{A}}. \tag{4.3}$$

Fig. 4.1.: Simulated data points in black and the corresponding trend filtering fit $\hat{a}_t$ in red.

Taking the ilr transform leads to the transformed problem

$$\min_{\mathrm{ilr}_{\mathbf{V}}(\boldsymbol{a}_t)\in\mathbb{R}^{D-1}} \sum_{t=1}^{T} \|\mathrm{ilr}_{\mathbf{V}}(\boldsymbol{x}_t) - \mathrm{ilr}_{\mathbf{V}}(\boldsymbol{a}_t)\|_E^2 + \lambda \sum_{t=3}^{T} \|\mathrm{ilr}_{\mathbf{V}}(\boldsymbol{a}_t) - 2\,\mathrm{ilr}_{\mathbf{V}}(\boldsymbol{a}_{t-1}) - \mathrm{ilr}_{\mathbf{V}}(\boldsymbol{a}_{t-2})\|_E. \tag{4.4}$$

In the latter a Euclidean penalty is put on the second differences of the ilr transformed quantities. This corresponds to a group lasso penalty and is known to have the same shrinkage properties on the second difference of $\mathrm{ilr}_{\mathbf{V}}(\boldsymbol{a}_t)$ than the l_1 penalty in the Lasso case, see [68]. Therefore, with growing λ, more $\boldsymbol{a}_t \ominus 2 \odot \boldsymbol{a}_{t-1} \oplus \boldsymbol{a}_{t-2}$ are exactly zero and so, as in the univariate case, if this holds for consecutive $t = t_0, \ldots, t_l$, then compositional linear trends are extracted. To solve problem (4.3) we developed a fast ADMM based algorithm, see [62]. Furthermore, we tested the utility of the method on data of COVID-19 infections in 9 different countries.

Publication III: Compositional trend filtering

Summary

Publication III considers an extension of univariate trend filtering of real valued time series to the compositional time series case. The trend filtering optimization problem is formulated in a compositional context and an efficient ADMM based algorithm for its solution is developed. The utility of the method is shown in an application to COVID-19 data.

Bibliographic information

C. Rieser, P. Filzmoser. Compositional trend filtering. Annales Mathematicae et Informaticae, 53, 257-270, 2021. https://doi.org/10.33039/ami.2021.02.004.

Author's contribution

C. Rieser developed the methodological part, including the optimization algorithm, implemented the method in R, wrote the first draft of the paper and its revision upon discussion with the co-author.

5. Functional outliers

In this chapter we recapture the main ideas for the detection of compositional functional outliers as developed in Publication IV. The main focus of this publication was to extend Euclidean based outlier detection for functional data to the compositional case.

5.1. Functional data

In functional data analysis (FDA) one considers samples that are whole functions, e.g. $\boldsymbol{f}_n : \Omega \to \mathbb{R}^D$, where Ω is typically a compact subset of $\mathbb{R}^p$. It is common to assume that each such data point is in the space of square integrable functions

$$\mathcal{L}^2(\Omega) := \left\{ \boldsymbol{f} : \Omega \to \mathbb{R}^D \middle| \sum_{i=1}^{D} \int_\Omega f_i(\boldsymbol{x})^2 d\boldsymbol{x} \right\}. \tag{5.1}$$

All coordinate functions f_i that are considered in this section have the same domain Ω, however, extensions to separate Ω_i are straightforward and are situation specific. Further extensions that allow different measures on each Ω_i are also possible. Nevertheless, to facilitate things we only describe the framework for functions with domain $\Omega = [a, b] \subset \mathbb{R}$ equipped with the standard Lebesgue measure. For two vector functions $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{L}^2(\Omega)$ an inner product and a corresponding norm are defined as

$$\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathcal{L}^2} := \sum_{i=1}^{D} \int_\Omega f_i(\boldsymbol{x}) g_i(\boldsymbol{x}) d\boldsymbol{x} \text{ and } \|\boldsymbol{f}\|_{\mathcal{L}^2} := \sqrt{\langle \boldsymbol{f}, \boldsymbol{g} \rangle_{\mathcal{L}^2}}.$$

As we are interested in $\boldsymbol{f}$ as a random element we additionally assume that we have a probability measure on $\mathcal{L}^2$. If for a random element $\boldsymbol{f}$, $\mathbb{E}(\|\boldsymbol{f}\|^2_{\mathcal{L}^2}) < \infty$ holds, then the mean of $\boldsymbol{f}$, denoted as $\mathbb{E}(\boldsymbol{f})$, exists and is defined as the unique element fulfilling

$$\mathbb{E}(\langle \boldsymbol{f}, \boldsymbol{u} \rangle_{\mathcal{L}^2}) = \langle \mathbb{E}(\boldsymbol{f}), \boldsymbol{u} \rangle_{\mathcal{L}^2},$$

for any $\boldsymbol{u} \in \mathcal{L}^2$. The covariance operator $\mathbf{C} : \mathcal{L}^2 \to \mathcal{L}^2$ is typically defined as

$$\boldsymbol{u} \mapsto \mathbb{E}\Big(\langle \boldsymbol{f} - \mathbb{E}(\boldsymbol{f}), \boldsymbol{u} \rangle_{\mathcal{L}^2} (\boldsymbol{f} - \mathbb{E}(\boldsymbol{f})) \Big) \tag{5.2}$$

and is the functional analogue of the covariance matrix for multivariate random vectors. The covariance operator $\mathbf{C}$ has certain attractive properties.

Proposition 4. *Let $\boldsymbol{u}, \boldsymbol{v}$ be any two elements in $\mathcal{L}^2$ then*

1) $\mathbf{C}$ *is symmetric, i.e. for $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{L}^2$ we have $\langle \mathbf{C}(\boldsymbol{u}), \boldsymbol{v} \rangle_{\mathcal{L}^2} = \langle \boldsymbol{u}, \mathbf{C}(\boldsymbol{v}) \rangle_{\mathcal{L}^2}$.*

2) $\mathbf{C}$ *is nonnegative-definite, $\langle \mathbf{C}(\boldsymbol{u}), \boldsymbol{u} \rangle_{\mathcal{L}^2} \geq 0$.*

3) $\mathbf{C}$ *has an eigen-decomposition,* $\mathbf{C}(\boldsymbol{u}) = \sum_{i=1}^{\infty} \lambda_i \langle \boldsymbol{u}_i, \boldsymbol{u} \rangle_{\mathcal{L}^2} \boldsymbol{u}_i$*, for an orthogonal basis of eigenvectors* $\boldsymbol{u}_i$ *to the eigenvalues* $\lambda_i \geq 0$*, with* $\sum_{i=1}^{\infty} \lambda_i < \infty$*.*

4) *The random element* $\boldsymbol{f}$ *can be written as*

$$\boldsymbol{f} = \mathbb{E}(\boldsymbol{f}) + \sum_{i=1}^{\infty} \zeta_i \boldsymbol{u}_i$$

where $\zeta_i := \langle \boldsymbol{u}_i, \boldsymbol{f} - \mathbb{E}(\boldsymbol{f}) \rangle_{\mathcal{L}^2}$*, with* $\mathbb{E}(\zeta_i) = 0$*,* $\mathbb{E}(\zeta_i^2) = \lambda_i$ *and* $\mathbb{E}(\zeta_i \zeta_j) = 0$*, for* $i \neq j$*.*

The decomposition in 4) is known as the Karhunen-Loeve expansion and is the functional pendant to PCA for multivariate data. For a thorough introduction to FDA we refer to [39] or [50]. Usually, functional data sets do not come in functional form but in the form of observed data for certain points in the domain. For each $n = 1, \ldots, N$, we have erroneous observations of $\boldsymbol{f}_n$ at times t_{nm}, with $t_{nm} \in \Omega$ and $m = 1, .., N_n$, given as $(t_{nm}, \boldsymbol{y}_{nm})$. From this data a functional data set is constructed. Depending on smoothness assumptions as well as assumptions on Ω different methods exist. In the case that Ω is an interval a typical approach is to solve for each n an optimization problem of the form

$$\hat{\boldsymbol{f}}_n := \arg\min \sum_{m=1}^{N_n} \left\| \boldsymbol{y}_{nm} - \boldsymbol{f}_n(t_{nm}) \right\|_E^2 + \lambda \int_{\Omega} \left\| \boldsymbol{f}_n''(t) \right\|_2^2 dt, \tag{5.3}$$

for a fixed smoothing parameter $\lambda > 0$. The functions $\boldsymbol{f}_n$ are usually modeled as multivariate splines, see [67], $\boldsymbol{f}_n(t) := \sum_{k=1}^{K} \boldsymbol{\beta}_k m_k(t)$, with $\boldsymbol{\beta}_k$ being coefficient vectors to be estimated and m_k some basis functions, such as B-splines or Fourier basis functions. Plugging this expansion into (5.3) leads to a convex optimization problem over the vectors $\boldsymbol{\beta}_k$ which can then be solved by standard methods. In case that observations at time points t_{nm} are scarce, other methods for constructing functional data $\boldsymbol{f}_n$, for example methods based on local polynomial kernel smoothing, exist, see [39]. As the Hilbert space framework is very general, many methods, such as linear regression, can readily be extended; see Equations (2.27)-(2.30).

5.2. Functional outliers

In general, an outlier can defined as an observation that is very different from the majority of the data. For example, in one dimension it is common to speak of outliers as points that are a certain (robust) deviation away from the (robust) center. In a multivariate setting the approach is similar. Given data $\boldsymbol{x}_n$, with $n = 1, \ldots, N$, a point $\boldsymbol{x}_{n_0}$ can be flagged as an outlier if its (robust) Mahalanobis distance, $\sqrt{(\boldsymbol{x}_{\boldsymbol{n_0}} - \boldsymbol{m})' \mathbf{C}^{-1} (\boldsymbol{x}_{\boldsymbol{n_0}} - \boldsymbol{m})}$, where $\boldsymbol{m}$ respectively $\mathbf{C}$ are robust estimates for the location respectively scatter, is above a certain threshold, see [53], [54] or [33]. Extending the Mahalanobis distance to functional data for the use of outlier detection is no easy task as already the non-robust pendant of the covariance matrix (5.2) is non-invertible, [39]. Some attempts in this direction have been made, [21] or [22], but most extensions of defining a distance measure that quantifies of how much outlying an observation is take a different approach by building on the concept of statistical depth. For example, in the case of univariate functional samples $\hat{f}_n : \Omega \to \mathbb{R}$, the Fraiman and Muniz

depth, see [20], of a functional point f is defined as

$$\text{depth}(f) := \int_\Omega \left|1 - \left|\frac{1}{2} - F_n(t)\right|\right| dt, \tag{5.4}$$

where $F_n(t)$ is the empirical distribution function of $\hat{f}_1(t), \ldots, \hat{f}_n(t)$ at time point t. Another example is the h-mode depth, see [10],

$$\text{depth}(f) = \frac{1}{Nh} \sum_{n=1}^{N} \mathcal{K}\left(\frac{\left\|f - \hat{f}_n\right\|}{h}\right), \tag{5.5}$$

for an appropriately chosen kernel $\mathcal{K}$ and tuning parameter $h > 0$. For multivariate functional data $\hat{\boldsymbol{f}}_n$ it seems natural to define a depth function pointwise, see [9], similar to (5.4):

$$\text{depth}(\boldsymbol{f}) := \int_\Omega \mathrm{D}(\boldsymbol{f}(t), P_n(t)) w(t) dt \tag{5.6}$$

where $P_n(t)$ is the empirical probability distribution of $\hat{\boldsymbol{f}}_n$ in t, $\mathrm{D} : \mathbb{R}^D \to \mathbb{R}_+$ is a depth function on $\mathbb{R}^D$ and $w : \Omega \to \mathbb{R}_+$ is some weighting function. Any depth function D on $\mathbb{R}^D$ needs to satisfy certain desirable properties, [69], of which the most important ones are invariance under affine transformations, maximality at a uniquely defined center, monotonicity relative to the center and vanishing at infinity. Functions $\boldsymbol{f}$ which are close to the center have a large depth score and functions far away have a low one. Note that with the aid of depth functions the concept of quantiles can also be generalized by ordering observations $\hat{\boldsymbol{f}}_n$ according to the ordered depth statistics. Commonly used depth statistics on $\mathbb{R}^D$ are, for a fixed $\boldsymbol{x}$, the halfspace depth

$$\mathrm{D}(\boldsymbol{x}, P) := \inf_{\|\boldsymbol{a}\|_E = 1} \mathrm{P}(\boldsymbol{y} | \boldsymbol{a}'\boldsymbol{y} \geq \boldsymbol{a}'\boldsymbol{x}), \tag{5.7}$$

where P is some probability measure on $\mathbb{R}^D$, and the projection depth

$$\mathrm{D}(\boldsymbol{x}, \mathrm{P}) := \left(1 + \sup_{\|\boldsymbol{a}\|_E = 1} \left(\frac{|\langle \boldsymbol{a}, \boldsymbol{x} \rangle_E - \text{med}(\langle \boldsymbol{a}, \boldsymbol{y} \rangle_E)|}{\text{mad}(\langle \boldsymbol{a}, \boldsymbol{y} \rangle_E)}\right)\right)^{-1} \tag{5.8}$$

where $\boldsymbol{y}$ is distributed as P, med is the median and mad the median absolute deviation, see [44] and [14].

5.3. Compositional functional outlier detection

In Paper IV we considered an extension of the functional depth (5.6) to the compositional case. In this paper we looked at compositional functional data $f_n : \Omega \to \mathcal{S}^D, n = 1, \ldots, N$. As in the non-compositional case one needs to estimate in a first step from the raw observations $(t_{nm}, \boldsymbol{y}_{nm})$, with $t_{nm} \in \Omega$, $m = 1, \ldots, N_n$ and $\boldsymbol{y}_{nm} \in \mathcal{S}^D$, the corresponding functional data. For this purpose, for each n the following optimization problem is solved

$$\hat{\boldsymbol{f}}_n := \arg\min \sum_{m=1}^{N_n} \left\|\boldsymbol{y}_{nm} \ominus \boldsymbol{f}_n(t_{nm})\right\|_{\mathcal{A}}^2 + \lambda \int_\Omega \left\|\boldsymbol{f}_n''(t)\right\|_{\mathcal{A}}^2 dt, \tag{5.9}$$

where the superscript $''$ of $\boldsymbol{f}_n(t)$ denotes the second compositional derivative, [47], and $\lambda > 0$ is a smoothing parameter controlling the smoothness of $\boldsymbol{f}_n$. Problem (5.9) can be solved for the ilr-transformed quantities through (5.3) and is invariant with respect to the choice of $\mathbf{V}$. In Paper IV we adapted the adjusted outlyingness to the compositional case,

$$AO(\boldsymbol{x}, \mathrm{P}) := \begin{cases} \sup_{\|\boldsymbol{a}\|=1} \left(\frac{\boldsymbol{a}'\boldsymbol{x} - \mathrm{med}(\boldsymbol{a}'\boldsymbol{y})}{w_2(\boldsymbol{a}'\boldsymbol{X}) - \mathrm{med}(\boldsymbol{a}'\boldsymbol{y})} \right) & \text{if } \boldsymbol{a}'\boldsymbol{x} > \mathrm{med}(\boldsymbol{a}'\boldsymbol{y}) \\ \sup_{\|\boldsymbol{a}\|=1} \left(\frac{\mathrm{med}(\boldsymbol{a}'\boldsymbol{y}) - \boldsymbol{a}'\boldsymbol{x}}{\mathrm{med}(\boldsymbol{a}'\boldsymbol{X}) - w_1(\boldsymbol{a}'\boldsymbol{X})} \right) & \text{if } \boldsymbol{a}'\boldsymbol{x} \leq \mathrm{med}(\boldsymbol{a}'\boldsymbol{y}), \end{cases} \tag{5.10}$$

where $\boldsymbol{y} \sim \mathrm{P}$ and w_1 and w_2 are functions that adjust for skewness in the projected samples, see [35] and [34]. The adjusted outlyingness is an extension of the projection depth (5.8) and was used in [34] to construct a measure of outlyingness of a real valued multivariate functional data point $\boldsymbol{f}$ as

$$\mathrm{depth}(\boldsymbol{f}) := \int_\Omega AO(\boldsymbol{f}(t), \mathrm{P}) w(t) dt,$$

where $w : \Omega \to \mathbb{R}_+$ is again some weighting function. Because (5.10) only contains Euclidean inner products in $\mathbb{R}^D$, adapting it to the compositional case corresponds to replacing the respective inner products with the compositional ones.

Publication IV: Outlier Detection for Pandemic-Related Data Using Compositional Functional Data Analysis

Summary

In Publication IV methods for detecting functional outliers were extended to the compositional setting. A method to construct compositional functional data points from raw observations, based on vector smoothing splines, is discussed. Measures for compositional functional outlyingness based on recently developed adjusted outlyingness for Euclidean functional data are developed. Graphical tools to analyse the results are also extended. The utility of the compositional approach as well as its difference to the Euclidean case is discussed with respect to COVID-19 data.

Bibliographic information

C. Rieser, P. Filzmoser. Outlier Detection for Pandemic-Related Data Using Compositional Functional Data Analysis. In: Boado-Penas M.C., Eisenberg J., Sahin S (eds). Pandemics: Insurance and Social Protection. Springer Actuarial, Springer, Cham, 2022, pp. 251-266. https://doi.org/10.1007/978-3-030-78334-1_12.

Author's contribution

C. Rieser developed the methodological part, implemented the method in R and wrote the first draft of the paper after discussion with the co-author.

6. Compositional data from a graph signal processing perspective

In this chapter we summarize the main ideas and concepts of Publication V. In the latter we looked at the connections of CoDA with Graph Signal Processing (GSP). The main idea was to modify the Aitchison geometry by constructing an inner product that, in contrast to the Aitchison inner product (2.9), does not penalize each log-ratio pair equally, but instead puts a different weighting on each. For two different vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}_+^D$ we investigated an extension of the Aitchison inner product as

$$\sum_{i,j=1}^{D} \log\left(\frac{x_i}{x_j}\right) \log\left(\frac{y_i}{y_j}\right) w_{ij}, \tag{6.1}$$

for some fixed symmetric and non-negative weights w_{ij}. Important properties of CoDA, such as scale invariance and subcompositional coherence, still hold. Furthermore, we constructed one-to-one mappings that allow us, as in the Aitchison setting, to map our data into the Euclidean space where usual methods based on the Euclidean geometry can be used. Different weighting approaches and investigations into the connection of CoDA and graph theory have appeared before, as for example in [30], [31], [25] and [26], but no connection to GSP was made, nor was a new (graph based) geometry investigated.

6.1. Graph signal processing

A graph $\mathcal{G}$ is a pair $(\mathcal{V}, \mathbf{W})$ with $\mathcal{V} = \{1, \ldots, D\}$ denoting a set of nodes and $\mathbf{W} = (w_{ij})_{1 \leq i,j, \leq D} \in \mathbb{R}^{D \times D}$ a symmetric matrix with non-negative elements and a zero diagonal. A graph can be used to model relationships between nodes as explained in the following. Assume for the moment that we have a real valued variable $\boldsymbol{f} \in \mathbb{R}^D$. Take for example a gray color image that is vectorized, that is the columns of the image are stacked into one vector. For an image it seems sensible to assume that the gray intensities f_i and f_j for the i-th and the j-th pixel are close, $f_i \approx f_j$, whenever the pixels i and j are. Setting weights w_{ij} to one when a pixel i is close to j and zero otherwise we can define a quantity

$$\frac{1}{2} \sum_{i,j=1}^{D} (f_i - f_j)^2 w_{ij}, \tag{6.2}$$

which measures how fast the intensities in neighboring pixels of the image $\boldsymbol{f}$ change for the overall image. After some refactoring, (6.2) above can be rewritten as

$$\frac{1}{2} \sum_{i,j=1}^{D} (f_i - f_j)^2 w_{ij} = \boldsymbol{f}' \mathbf{L}_{\mathbf{W}} \boldsymbol{f}, \tag{6.3}$$

see [42] for a proof, where $\mathbf{L_W} = \text{diag}(\mathbf{W1}) - \mathbf{W} \in \mathbb{R}^{D\times D}$ is the so-called Laplacian matrix. In fact, for two real valued vectors $\boldsymbol{f}$ and $\boldsymbol{g}$ a more general result

$$\frac{1}{2}\sum_{i,j=1}^{D}(f_i - f_j)(g_i - g_j)w_{ij} = \boldsymbol{f}'\mathbf{L_W}\boldsymbol{g} \tag{6.4}$$

also holds. The Laplacian matrix $\mathbf{L_W}$ has two important properties:

- $\mathbf{L_W}$ is a symmetric positive semi-definite matrix.
- The constant vector $\mathbf{1}$ is always in the null space of $\mathbf{L_W}$, i.e. $\mathbf{L_W 1} = \mathbf{0}$.

For a proof and more on Laplacian matrices see [42] or [43]. The eigenvectors of the Laplacian matrix are of special importance in GSP. Computing the eigenvector corresponding to the k-th smallest eigenvalue is equivalent to minimizing the left-hand side of (6.3) over the vector $\boldsymbol{f} \in \mathbb{R}^D$ under additional constraints :

$$\min_{\|\boldsymbol{f}\|_E=1} \frac{1}{2}\sum_{i,j=1}^{D}(f_i - f_j)^2 w_{ij} \text{ such that } \langle \boldsymbol{f}, \boldsymbol{f}^l\rangle_E = 0 \text{ for } l = 1,\dots,k-1, \tag{6.5}$$

where $\boldsymbol{f}^l$ are the eigenvectors to the l-th smallest eigenvalue. As pairs of entries (f_i, f_j) of a solution $\boldsymbol{f}$ of (6.5) are forced to be similar for adjacent nodes, that is nodes $i, j \in \{1,\dots,D\}$ for which $w_{ij} \neq 0$ holds, $\boldsymbol{f}$ can be interpreted as a function that varies comparatively little over connected nodes. If the graph $\mathcal{G}$ is connected, that is from every node $i \in \mathcal{V}$ to any other $j \in \mathcal{V}$ we can find a sequence of indices $i = 1_1, i_2, \dots, i_m = j$ with positive weights $w_{i_s i_{s+1}} \neq 0$, $s = 1,\dots,m-1$, then $\frac{1}{\sqrt{D}}\mathbf{1}$ is the only eigenvector to the zero eigenvalue. A prominent example is the line graph. For the latter, which is defined as the graph with weights $w_{12} = w_{23} = \dots = w_{(D-1)D} = 1$ and zero otherwise, the eigenvectors of $\mathbf{L_W}$ are, besides the trivial eigenvector $\frac{1}{\sqrt{D}}\mathbf{1}$, equal to

$$(\boldsymbol{f}^l)_i = \sqrt{\frac{2}{D}}\cos\left(\frac{\pi(l-1)(i-0.5)}{D}\right)$$

for $l = 2,\dots,D$, see [56]. These eigenvectors are smooth with respect to i and can be thought of as a discretized version of squashed cosine functions on $\mathbb{R}$. Similar reasoning can be applied to $\mathbf{L_W}$ being the discretized version of the Laplace operator $\boldsymbol{\Delta}$ in Analysis for a grid graph, see [24]. Another example is the ring graph whose eigenvectors lead to the discrete Fourier transform. For all these reasons one can think of the eigenvectors of $\mathbf{L_W}$ as in (6.5) as a graph analogue to the Fourier basis functions. Continuing the analogy, the Graph Fourier Transform of a signal $\boldsymbol{g} \in \mathbb{R}^D$ is defined as the vector

$$(\hat{g})_{l=2,\dots,D} := (\langle \boldsymbol{g}, \boldsymbol{f}^l\rangle_E)_{l=2,\dots,D},$$

see [56]. Through these analogies it is also possible to define in a sensible way a convolution operation on graphs, see [56] and other extensions, such as wavelets on graphs, see [27].

6.2. CoDA in a GSP context

The left-hand side of (6.4) is, for the log-transformed variables $\boldsymbol{f} := \log(\boldsymbol{x})$ and $\boldsymbol{g} := \log(\boldsymbol{y})$, equal to (6.1). From this and the fact that $\mathbf{1}$ is in the kernel of $\mathbf{L}_\mathbf{W}$ follows that rescaling elements $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}_+^D$ does not change (6.1). Therefore, in Publication V we investigated a generalization to the Aitchison geometry building on the semi-inner product

$$\sum_{i,j=1}^{D} \log\left(\frac{x_i}{x_j}\right) \log\left(\frac{y_i}{y_j}\right) w_{ij} = \log(\boldsymbol{x})' \mathbf{L}_\mathbf{W} \log(\boldsymbol{y}). \tag{6.6}$$

Similar to Section 2.4 one can construct an equivalence relation as

$$\log(\boldsymbol{x}) \sim \log(\boldsymbol{y}) \iff \mathbf{L}_\mathbf{W}(\log(\boldsymbol{x}) - \log(\boldsymbol{y})) = \mathbf{0} \tag{6.7}$$

such that (6.7) becomes an inner product on the set of equivalence classes $\{[\log(\boldsymbol{x})] | \boldsymbol{x} \in \mathbb{R}_+^D\}$ with $[\log(\boldsymbol{x})] := \{\log(\boldsymbol{y}) | \log(\boldsymbol{y}) \sim \log(\boldsymbol{x})\}$. A more convenient way is to fix the representatives for each class $[\log(\boldsymbol{x})]$, again similar to Section 2.4. In Paper V we defined in parallel to the D-part simplex, the D-part graph Simplex as

$$\mathcal{S}_\mathbf{W}^D := \left\{ \boldsymbol{x} \in \mathbb{R}_+^D \,\middle|\, \sum_{i \in \mathcal{V}_m} x_i = 1, m = 1, \ldots, M \right\} \tag{6.8}$$

where $\mathcal{V}_m$ are non-overlapping connected subsets of nodes that form a partition of $\mathcal{V}$. As in the Aitchison geometry, one can show that with slightly adapted perturbation and powering operation this space is again a Hilbert space. A natural extension of the clr-map as well as one-to-one isometries, bijective maps which are angle and distance preserving, also exist, and the latter allow us to map data into a lower dimensional Euclidean space $\mathbb{R}^{D-M}$. One natural one-to-one isometry follows from the eigen-decomposition of $\mathbf{L}_\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}'$, where $\mathbf{U} \in \mathbb{R}^{D \times D}$ is a unitary matrix and $\mathbf{\Sigma}$ is a diagonal matrix with non-negative entries, ordered from highest to lowest eigenvalue, of which M are exactly zero. In Publication V it is shown that the map

$$\boldsymbol{x} \mapsto \sqrt{\Sigma_{ii}} \langle \boldsymbol{u}_i, \log(\boldsymbol{x}) \rangle_E, \tag{6.9}$$

with Σ_{ii} being the diagonal elements of $\mathbf{\Sigma}$ and $\boldsymbol{u}_i$ the i-th column of $\mathbf{U}$, $i = 1, \ldots, D - M$, is one-to-one and also fulfills for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}_\mathbf{W}^D$

$$\log(\boldsymbol{x})' \mathbf{L}_\mathbf{W} \log(\boldsymbol{y}) = \sum_{i=1}^{D-M} \Sigma_{ii} \langle \boldsymbol{u}_i, \log(\boldsymbol{x}) \rangle_E \langle \boldsymbol{u}_i, \log(\boldsymbol{y}) \rangle_E.$$

This property shows that (6.9) is isometric to $\mathbb{R}^{D-M}$ and therefore has the same role than the ilr map in CoDA. Further isometries as well as interpretable maps are also investigated.

Publication V: Extending compositional data analysis from a graph signal processing perspective

Summary

In Publication V we show that CoDA can be put into the grander framework of signal processing on graphs and investigate their link between these two different fields. Major concepts from CoDA are extend without loss of scale invariance or subcompositional coherence. An extensive theory is developed and two applications are discussed.

Bibliographic information

C. Rieser, P. Filzmoser. Extending compositional data analysis from a graph signal processing perspective. Submitted for publication.

Author's contribution

C. Rieser developed the methodological part, implemented the method in R, and wrote the paper upon extensive discussions with the co-author.

7. Conclusions

Research in CoDA has been and continues to be very active since the core principles were laid out in groundbreaking works by Aitchison [1], [3] and [2]. Recently, many ideas of CoDA have been adapted to new areas where the mathematical objects of interest are more appropriately dealt with as being parts of a whole. One such area is the statistical analysis of continuous density data, see [17], [60], [32], [57] and [41], which can be seen as an infinite extension of elements of the D-part simplex. Adapted tools from CoDA allows handling continuous density data in a very elegant way. In general, treating data as compositional can, but not always will, lead to quite different results. As we are more used to the Euclidean way of thinking, it is important to carefully adapt regular Euclidean methods when putting them into a CoDA framework. Multivariate positive data are omnipresent, and quite often the coordinates should be considered as part of a whole. However, in many areas such as biology, for example, CoDA is still not used as often as it should be, even though it has long been established that the compositional way is the more appropriate one, see [40]. Nevertheless, as the body of methods of CoDA grows, so does the amount of applications spanning over a big number of different areas and helping to shed light onto data from a different angle. This applicability in many different areas is also reflected in this thesis, with publications ranging from methodologies developed in the geochemical sciences, over multivariate time series, functional outliers with application to COVID-19 related data, to more theoretical and emerging areas such as graph signal processing. Many more extensions can still be considered in a lot of different areas. In this thesis we looked only at a very small subset, and further research could be carried out for each publication. For example, as in Publication V the graph Laplacian matrix is of central importance, it would be very interesting to investigate further different methods of finding the weights and looking into their implications. Additionally, it would be interesting to investigate these estimators then under different types of outliers. Another direction could be pursued for compositional functional data and corresponding outliers, as investigated in Publication IV. In the latter we investigated a novel version of compositional functional depth to detect outliers. However, we only considered the case where the image of each functional data point is compositional and the domain is an interval in $\mathbb{R}$. It might also be interesting to consider functional data with domain also in the D-part simplex. Other areas that are also worth investigating upon are generalized additive models and their interplay with compositional regressors. All in all, we can say that there are still many open questions which are worth investigating.

References

1. Aitchison J. A new approach to null correlations of proportions. *Journal of the International Association for Mathematical Geology* **13,** 175–189 (1981).
2. Aitchison J. Principal component analysis of compositional data. *Biometrika* **70,** 57–65 (1983).
3. Aitchison J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)* **44,** 139–177 (1982).
4. Aitchison J. *The Statistical Analysis of Compositional Data* (Chapman & Hall, London, 1986).
5. Akhiezer NI & Glazman IM. *Theory of Linear Operators in Hilbert space* (Courier Corporation, 2013).
6. Arnold TB & Tibshirani RJ. *genlasso: Path Algorithm for Generalized Lasso Problems* R package version 1.5 (2020).
7. Barceló-Vidal C, Martín-Fernández JA & Pawlowsky-Glahn V. *Mathematical foundations of compositional data analysis* in *Proceedings of the Sixth Annual Conference of the International Association for Mathematical Geology* (ed Ross G) (2001), 1–20.
8. Chen J, Zhang X & Li S. Multiple linear regression with compositional response and covariates. *Journal of Applied Statistics* **44,** 2270–2285 (2017).
9. Claeskens G, Hubert M, Slaets L & Vakili K. Multivariate functional halfspace depth. *Journal of the American Statistical Association* **109,** 411–423 (2014).
10. Cuevas A, Febrero M & Fraiman R. On the use of the bootstrap for estimating functions with functional data. *Computational Statistics & Data Analysis* **51,** 1063–1074 (2006).
11. Debnath L & Mikusinski P. *Introduction to Hilbert Spaces with Applications* (Academic press, 2005).
12. Do Carmo MP. *Differential Geometry of Curves and Surfaces: Revised and Updated Second Edition* (Courier Dover Publications, 2016).
13. Dobson AJ & Barnett AG. *An Introduction to Generalized Linear Models* (CRC press, 2018).
14. Donoho DL, Gasko M, *et al.* Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics* **20,** 1803–1827 (1992).
15. Duchon J. *Splines minimizing rotation-invariant semi-norms in Sobolev spaces* in *Constructive Theory of Functions of Several Variables* (eds Schempp W & Zeller K) (Springer Berlin Heidelberg, Berlin, Heidelberg, 1977), 85–100.

16. Dumuid D *et al.* The compositional isotemporal substitution model: a method for estimating changes in a health outcome for reallocation of time between sleep, physical activity and sedentary behaviour. *Statistical Methods in Medical Research* **28,** 846–857 (2019).

17. Egozcue JJ, Díaz-Barrero JL & Pawlowsky-Glahn V. Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica* **22,** 1175–1182 (2006).

18. Egozcue JJ *et al.* Elements of simplicial linear algebra and geometry. *Compositional Data Analysis: Theory and Applications,* 141–157 (2011).

19. Filzmoser P, Hron K & Templ M. *Applied Compositional Data Analysis: With Worked Examples in R* (Springer, 2018).

20. Fraiman R & Muniz G. Trimmed means for functional data. *Test* **10,** 419–440 (2001).

21. Galeano P, Joseph E & Lillo RE. The Mahalanobis distance for functional data with applications to classification. *Technometrics* **57,** 281–291 (2015).

22. Ghiglietti A, Ieva F & Paganoni AM. Statistical inference for stochastic processes: two-sample hypothesis tests. *Journal of Statistical Planning and Inference* **180,** 49–68 (2017).

23. Gloor GB, Macklaim JM, Pawlowsky-Glahn V & Egozcue JJ. Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology* **8,** 2224 (2017).

24. Grady LJ & Polimeni JR. *Discrete Calculus: Applied Analysis on Graphs for Computational Science* (Springer, 2010).

25. Greenacre M. Variable selection in compositional data analysis using pairwise logratios. *Mathematical Geosciences* **51,** 649–682 (2019).

26. Greenacre M, Grunsky E & Bacon-Shone J. A comparison of isometric and amalgamation logratio balances in compositional data analysis. *Computers & Geosciences* **148,** 104621 (2021).

27. Hammond DK, Vandergheynst P & Gribonval R. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis* **30,** 129–150 (2011).

28. Hastie T, Tibshirani R & Friedman JH. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, 2009).

29. Hodrick RJ & Prescott EC. Postwar US business cycles: an empirical investigation. *Journal of Money, Credit, and Banking,* 1–16 (1997).

30. Hron K, Engle M, Filzmoser P & Fišerová E. Weighted symmetric pivot coordinates for compositional data with geochemical applications. *Mathematical Geosciences* **53,** 655–674 (2021).

31. Hron K, Filzmoser P, de Caritat P, Fišerová E & Gardlo A. Weighted pivot coordinates for compositional data and their application to geochemical mapping. *Mathematical Geosciences* **49,** 797–814 (2017).

32. Hron K, Menafoglio A, Templ M, Hrzová K & Filzmoser P. Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis* **94,** 330–350 (2016).

33. Hubert M & Debruyne M. Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics* **2,** 36–43 (2010).

34. Hubert M, Rousseeuw PJ & Segaert P. Multivariate functional outlier detection. *Statistical Methods & Applications* **24,** 177–202 (2015).

35. Hubert M & Vandervieren E. An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis* **52,** 5186–5201 (2008).

36. Jorgensen B. *The Theory of Dispersion Models* (CRC Press, 1997).

37. Kim SJ, Koh K, Boyd S & Gorinevsky D. l_1 trend filtering. *SIAM Review* **51,** 339–360 (2009).

38. Kline M. *Calculus: An Intuitive and Physical Approach* (Courier Corporation, 1998).

39. Kokoszka P & Reimherr M. *Introduction to Functional Data Analysis* (CRC Press, 2017).

40. Lovell D, Müller W, Taylor J, Zwart A & Helliwell C. *Proportions, percentages, ppm: do the molecular biosciences treat compositional data right?* in *Compositional Data Analysis: Theory and Applications* (eds Pawlowsky-Glahn V & Buccianti A) (London: John Wiley & Sons, Ltd., 2011), 193 –207.

41. Machalova J, Hron K & Monti GS. Preprocessing of centred logratio transformed density functions using smoothing splines. *Journal of Applied Statistics* **43,** 1419–1435 (2016).

42. Merris R. Laplacian matrices of graphs: a survey. *Linear Algebra and its Applications* **197,** 143–176 (1994).

43. Mohar B. *The Laplacian spectrum of graphs* in *Graph Theory, Combinatorics, and Applications* **2** (Wiley, 1991), 871–898.

44. Mosler K. *Depth statistics* in *Robustness and Complex Data Structures, Festschrift in Honour of Ursula Gather* (eds Becker C, Fried R & Kuhnt S) (Springer, 2013), 17–34.

45. Nelder JA & Wedderburn RW. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* **135,** 370–384 (1972).

46. Papritz A & Stein A. *Spatial prediction by linear kriging* in *Spatial Statistics for Remote Sensing* (eds Stein A, van der Meer F & Gorte B) (Springer, 1999), 83–113.

47. Pawlowsky-Glahn V & Buccianti A. *Compositional Data Analysis: Theory and Applications* (John Wiley & Sons, 2011).

48. Pearson K. Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London* **60,** 489–498 (1897).

49. Quinn TP *et al.* A field guide for the compositional analysis of any-omics data. *GigaScience* **8,** giz107 (2019).

50. Ramsay JO. Functional data analysis. *Encyclopedia of Statistical Sciences* **4** (2004).

51. Ravel J *et al.* Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences USA* **108,** 4680–4687 (2011).

52. Roman S. *Advanced Linear Algebra* 2nd ed. (Springer, 2005).

53. Rousseeuw PJ. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications* **8,** 37 (1985).

54. Rousseeuw PJ & Driessen KV. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41,** 212–223 (1999).

55. Sadhanala V & Tibshirani RJ. Additive models with trend filtering. *arXiv:1702.05037* (2017).

56. Shuman DI, Ricaud B & Vandergheynst P. Vertex-frequency analysis on graphs. *Applied and Computational Harmonic Analysis* **40,** 260–291 (2016).

57. Talská R, Menafoglio A, Machalová J, Hron K & Fišerová E. Compositional regression with functional response. *Computational Statistics & Data Analysis* **123,** 66–85 (2018).

58. Templ M, Hron K & Filzmoser P. *robCompositions: an R-package for robust statistical analysis of compositional data* in *Compositional Data Analysis: Theory and Applications* (eds Pawlowsky-Glahn V & Buccianti A) (London: John Wiley & Sons, Ltd., 2011), 341–355.

59. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58,** 267–288 (1996).

60. Van den Boogaart K, Egozcue J & Pawlowsky-Glahn V. Bayes Hilbert spaces. *Australian & New Zealand Journal of Statistics* **56,** 171–194 (2014).

61. Varma R, Lee H, Kovačević J & Chi Y. Vector-valued graph trend filtering with non-convex penalties. *IEEE Transactions on Signal and Information Processing over Networks* **6,** 48–62 (2019).

62. Wahlberg B, Boyd S, Annergren M & Wang Y. An ADMM algorithm for a class of total variation regularized estimation problems. *IFAC Proceedings Volumes* **45,** 83–88 (2012).

63. Wang YX, Sharpnack J, Smola AJ & Tibshirani RJ. Trend filtering on graphs. *The Journal of Machine Learning Research* **17,** 3651–3691 (2016).

64. Wood SN. *Generalized Additive Models: An Introduction with R* (CRC press, 2017).

65. Wood SN. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* **62,** 1025–1036 (2006).

66. Wood SN, Bravington MV & Hedley SL. Soap film smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70,** 931–955 (2008).

67. Yee TW. *Vector Generalized Linear and Additive Models: With an Implementation in R* (Springer, 2015).

68. Yuan M & Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68,** 49–67 (2006).

69. Zuo Y & Serfling R. General notions of statistical depth function. *Annals of Statistics,* 461–482 (2000).

Christopher Rieser

Date of Birth: 19.12.1988
Nationality: Austrian
Residence: Vienna
Phone: [redacted]

christopher.rieser@tuwien.ac.at

Favorite quote

"Be the change that you wish to see in the world." - M.Gandhi

Languages

* German ●●●●●
* English ●●●●●
* French ●●●●○
* Japanese ●●○○○

Skills

* R ●●●●○
* Matlab ●●●○○

Interests

* Parkour
* Calisthenics
* Sailing
* Reading

Current position

2018 – **FWF project assistant - CSTAT group of TU Wien**

Education

2018 – **Doctoral program**

○ Doctoral program in Natural Sciences Technical Mathematics, TU Wien, Austria.

2017 – 2018 **Dipl.-Ing. in Statistics**

○ Master program in Technical Mathematics, TU Wien, Austria. Specialization: Statistics.

2013 – 2016 **MSc in Mathematics**

○ Master program in Mathematics (Analysis), University of Vienna, Austria. Specialization: Analysis.

2009 – 2013 **BSc in Mathematics**

○ Bachelor program in Technical Mathematics, TU Graz, Austria

2003 – 2008 **Diplôme du Baccalauréat**

○ Option Internationale du Baccalauréat at the Lycée International, France. Specialization: ES

Grants

2020 **FWF Grant**

Austrian Science Fund grant of approximately 130.000 €

Publications

C. Rieser, P. Filzmoser. Extending compositional data analysis from a graph signal processing perspective. Submitted for publication.

C. Rieser, P. Filzmoser. Outlier Detection for Pandemic-Related Data Using Compositional Functional Data Analysis. In: Boado-Penas M.C., Eisenberg J., Sahin S (eds). Pandemics: Insurance and Social Protection. Springer Actuarial, Springer, Cham, 2022. https://doi.org/10.1007/978-3-030-78334-1_12.

C. Rieser, P. Filzmoser. Compositional trend filtering. Annales Mathematicae et Informaticae, 53, 257-270, 2021. https://doi.org/10.33039/ami.2021.02.004.

Christopher Rieser

Date of Birth: 19.12.1988
Nationality: Austrian
Residence: Vienna
Phone: [redacted]

christopher.rieser@tuwien.ac.at

Favorite quote

"Be the change that you wish to see in the world." - M.Gandhi

Languages

* German ●●●●●
* English ●●●●●
* French ●●●●○
* Japanese ●●○○○

Skills

* R ●●●●○
* Matlab ●●●○○

Interests

* Parkour
* Calisthenics
* Sailing
* Reading

Publications

D. Mikšová, C. Rieser, P. Filzmoser, M. Middleton, and R. Sutinen. Identification of mineralization in geochemistry for grid sampling using Generalized Additive Models. Mathematical Geosciences 53, 1861–1880, 2021. https://doi.org/10.1007/s11004-021-09929-x.

D. Mikšová, C. Rieser, and P. Filzmoser. Identification of mineralization in geochemistry along a transect based on the spatial curvature of log-ratios. Mathematical Geosciences 53, 1513-1533, 2021. https://doi.org/10.1007/s11004-021-09930-4.

Presentations

2022

- *Talk*: "Compositional Data and graph theory" at Innovation and Society 5.0: Statistical and Economic Methodologies for Quality Assessment (IES 2022), Capua, Italy.

2021

- *Talk*: "Multivariate functional outlier detection for compositions" at Online Conference Data Science, Statistics & Visualization (DSSV 2021), Rotterdam, The Netherlands.

2020

- *Talk*: "Compositional Trend Filtering" at Online Conference - 1st Conference on Information Technology and Data Science (CITDS 2020), Debrecen, Hungary.

2019

- *Talk*: "Piecewise smoothing splines" at Austrian Statistical Days (ÖSG Statistiktage 2019), Vienna, Austria.
- *Poster*: "Detection of mineralization using the second derivative of log-ratios" at 8th international Workshop on Compositional Data (CoDaWork), Terrassa, Spain.

Attended conferences

2022

- Innovation and Society 5.0: Statistical and Economic Methodologies for Quality Assessment (IES 2022), Capua, Italy.

2021

- International Conference on Robust Statistics (ICORS 2021), Vienna, Austria.
- Data Science, Statistics & Visualization (DSSV 2021), Rotterdam, The Netherlands.

Christopher Rieser

Date of Birth: 19.12.1988
Nationality: Austrian
Residence: Vienna
Phone: [redacted]

christopher.rieser@tuwien.ac.at

Favorite quote

"Be the change that you wish to see in the world." - M.Gandhi

Languages

* German ●●●●●
* English ●●●●●
* French ●●●●○
* Japanese ●●○○○

Skills

* R ●●●●○
* Matlab ●●●○○

Interests

* Parkour
* Calisthenics
* Sailing
* Reading

Attended conferences

2020

- Information Technology and Data Science (CITDS 2020), Debrecen, Hungary.

2019

- 8th international Workshop on Compositional Data (CoDaWork), Terrassa, Spain
- Austrian Statisitcal Days, Vienna, Austria.

2018

- Data Science, Statistics & Visualisation (DSSV) - Vienna, Austria
- Mountain Village Science Series (MOVISS), Vorau, Austria.